CEA-LIST at CheckThat! 2025: Evaluating LLMs as Detectors of Bias and Opinion in Text

Notebook for the CheckThat! Lab at CLEF 2025

Akram Elbouanani[1], Evan Dufraisse[1], Aboubacar Tuo[1] and Adrian Popescu[1]

[1]Université Paris-Saclay, CEA-List, F-91120, Palaiseau, France

Abstract

This paper presents a competitive approach to multilingual subjectivity detection using large language models (LLMs) with few-shot prompting. We participated in Task 1: Subjectivity of the CheckThat! 2025 evaluation campaign. We show that LLMs, when paired with carefully designed prompts, can match or outperform fine-tuned smaller language models (SLMs), particularly in noisy or low-quality data settings. Despite experimenting with advanced prompt engineering techniques—such as debating LLMs and various example selection strategies—we found limited benefit beyond well-crafted standard few-shot prompts. Our system achieved top rankings across multiple languages in the CheckThat! 2025 subjectivity detection task, including first place in Arabic and Polish, and top-four finishes in Italian, English, German, and multilingual tracks. Notably, our method proved especially robust on the Arabic dataset, likely due to its resilience to annotation inconsistencies. These findings highlight the effectiveness and adaptability of LLM-based few-shot learning for multilingual sentiment tasks, offering a strong alternative to traditional fine-tuning, particularly when labeled data is scarce or inconsistent.

Keywords

Large Language Models, Few-Shot Learning, Prompt Engineering, Subjectivity Detection, Debate Prompting, Multilingual NLP

1. Introduction

In natural language processing, distinguishing between subjective and objective language is a foundational task with wide-ranging implications. Subjectivity refers to expressions that convey personal opinions, beliefs, sentiments, or evaluations, whereas objectivity is characterized by verifiable facts and observable phenomena independent of individual interpretation [1]. Accurate subjectivity detection plays a crucial role in applications such as fact-checking, media analysis, and content moderation, where separating opinion from factual reporting is essential to counter misinformation and uphold media integrity [2]. In legal domains, detecting subjective language in witness testimonies or contractual texts can inform interpretation and decision-making [3]. Similarly, in sentiment analysis, identifying subjective expressions enables a deeper understanding of public attitudes toward products, services, or events [3, 4].

The emergence of Large Language Models (LLMs) has presented new opportunities for text analysis tasks. Unlike traditional Smaller Language Models (SLMs), which often require extensive, task-specific fine-tuning on large annotated datasets, LLMs leverage their pre-training to address numerous tasks with minimal, or even no, additional training [5]. Their ability to recognize subtle linguistic cues, such as irony or sarcasm, makes them particularly suited for identifying subjective expressions. Furthermore, their adaptability enables them to conform more precisely to specific annotation guidelines, offering flexibility in defining and identifying subjective content.

In previous editions of the CheckThat! lab, the application of LLMs for tasks such as subjectivity detection has been less prevalent compared to fine-tuned SLMs [6, 7]. In this context, in the rare instances where they were used, their performance has not consistently exceeded that of optimized SLMs [8]. This

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

elbouanani.akram@gmail.com (A. Elbouanani); evan.dufraisse@cea.fr (E. Dufraisse); aboubacar.tuo@cea.fr (A. Tuo); adrian.popescu@cea.fr (A. Popescu)

0009-0005-0597-7149 (A. Elbouanani)

work aims to demonstrate that LLMs can indeed compete with, and potentially outperform, fine-tuned SLMs in subjectivity detection tasks through the application of techniques such as careful prompting and few-shot prompting. We investigate the extent to which precise prompt engineering and the provision of relevant examples within the prompt can enhance the performance of LLMs, thereby showcasing their potential for robust and adaptable subjectivity detection in academic and real-world settings. While these techniques show promise in theory, our experiments reveal that they do not consistently improve performance on the tested datasets, suggesting important limitations and directions for further research.

2. Related Work

The computational detection of subjective language has evolved significantly from its early rule-based foundations to contemporary data-driven approaches. This progression reflects broader trends in natural language processing while addressing challenges in identifying opinionated content. Current research emphasizes two critical aspects: (1) the development of increasingly sophisticated models capable of capturing linguistic nuance [9, 10], and (2) the creation of specialized techniques to optimize these models for subjectivity analysis [11, 12, 13].

Architectural Evolution. Traditional lexicon-based and supervised learning approaches have given way to transformer-based models, with BERT-style architectures demonstrating strong performance on binary subjectivity classification [14, 15, 16]. However, the emergence of LLMs has introduced new capabilities in detecting implicit subjectivity through contextual reasoning. Indeed, a key advantage of LLMs is their advanced ability to recognize subtle linguistic cues like irony, sarcasm, or implicit bias [17]. Their architectural design enables them to capture complex dependencies within sequential data, leading to a deeper understanding of intricate relationships between words and emotions. Recent empirical studies have demonstrated that LLMs consistently achieve higher overall accuracy in sentiment analysis, often outperforming specialized pre-trained transformer models due to their comprehensive grasp of human thought and emotion [10]. This ability indicates strong potential for subjectivity detection, as recognizing nuanced evaluative language is key to distinguishing subjective from objective statements [1].

Prompting Strategies. Despite their inherent capabilities, optimizing LLM performance in specialized tasks like subjectivity detection often requires appropriate prompting. While a wide range of prompting strategies exists, we focus in this work on a selected subset of strategies which we evaluate in our experiments.

- **Prompt engineering** involves meticulously designing input queries to guide LLMs toward desired outputs. This technique is crucial for clearly defining the nuances of subjectivity for the model, specifying output formats, and aligning the LLM's reasoning with task-specific objectives [18, 19]. However, prompt effectiveness can be highly sensitive to wording, and small changes may lead to inconsistent results [20, 21].
- **In-context Learning (ICL)** refers to the ability of LLMs to perform tasks by conditioning on input-output examples provided directly in the prompt, without updating model parameters. A common subcategory of ICL is **few-shot learning**, where the prompt includes a small number of labeled examples to help the model infer the task and classification criteria [22, 5]. A key challenge in few-shot ICL is selecting representative and diverse examples, as LLMs can overfit to or ignore suboptimal demonstrations.
- **Multi-agent LLM Systems** is an emerging paradigm to enhance LLM performance. This approach distributes responsibilities across multiple specialized agents, each focusing on specific functions like information retrieval, complex reasoning, or decision-making [23]. Multi-agent systems offer several advantages, including enhanced reliability through cross-verification, refined decision-making through collaborative information sharing, and improved handling of complex tasks by dividing workloads [23, 24, 25]. Yet, coordination overhead and potential inconsistencies between agents remain significant challenges [23].

The CheckThat! lab [26, 27], organized within CLEF, serves as a prominent platform for advancing subjectivity detection research, particularly in distinguishing subjective from objective statements at the sentence level within news articles across multiples languages [7]. CheckThat! evaluations have systematically demonstrated the strengths and limitations of different approaches to fact-checking and subjectivity detection. In early iterations (2018-2020), traditional machine learning models, such as SVM with carefully engineered linguistic features, achieved competitive results, particularly for English texts [28, 29]. The 2020-2021 evaluations marked a transition period in which fine-tuned BERT-style models began to dominate the leaderboards [30, 31]. These results established transformer architectures as the new baseline for subjectivity detection tasks. The most recent CheckThat! cycles (2023-2024) included early approaches relying on LLMs. While early submissions underperformed due to inadequate prompting strategies, subsequent systems demonstrated that properly optimized LLMs could match or exceed specialized models [6, 7].

3. Dataset

We utilize the provided multilingual dataset, which comprises sentence-level annotations labeled as either OBJ (objective) or SUBJ (subjective). Table 1 summarizes the number of annotated sentences per language and split. The dataset exhibits class imbalance across languages and splits, with some languages (e.g., Italian and Arabic) showing a predominance of OBJ labels, while others (e.g., Bulgarian) present a more balanced distribution.

During the exploratory phase, we focus primarily on English and Arabic. These two languages were selected due to their differing class distributions and dataset sizes. We operate under the assumption that insights obtained from these languages are transferable to the other languages in the dataset.

Table 1
Number of sentences and class distribution (OBJ vs SUBJ) across languages and data splits.

Language	Split	Sentences	OBJ	SUBJ
English	Train	830	532	298
	Dev	462	222	240
	Dev-Test	484	362	122
Italian	Train	1613	1231	382
	Dev	667	490	177
	Dev-Test	513	377	136
German	Train	800	492	308
	Dev	491	317	174
	Dev-Test	337	226	111
Bulgarian	Train	729	406	323
	Dev	467	175	139
	Dev-Test	250	143	107
Arabic	Train	2446	1391	1055
	Dev	742	266	201
	Dev-Test	748	425	323

4. Methodology

We explore several strategies to improve subjective sentence classification. We experiment with the following three main approaches: prompt engineering, few-shot learning, and multi-agent LLM setups. We also use fine-tuned SLMs as baselines.

4.1. Prompt Engineering

We systematically evaluate the impact of prompt phrasing and label framing on classification performance. We compare:

- A minimal, generic prompt vs. a detailed one generated from the annotation guidelines.
- Label framing using explicit terms ("Subjective"/"Objective") vs. neutral terms ("Category 0"/"Category 1").
- Binary yes/no questions (e.g., "Is the sentence subjective?") as an alternative to direct classification.

These variations are designed to probe how linguistic framing affects the model's interpretability and consistency.

4.2. Few-Shot Learning Strategies

We experiment with a range of few-shot prompting configurations to assess how the number and selection of support examples influence performance. Specifically, we compare:

- Prompting setups with 0 (no-shot), 6-shot, and 12-shot examples.
- Example selection strategies based on: (a) Semantic similarity, (b) Semantic dissimilarity, and (c) Random sampling.

For the semantic approaches, similarity and dissimilarity are measured using cosine similarity between sentence embeddings generated by OpenAI's *text-embedding-3-small* model. To ensure class balance in the few-shot prompts, we selected an equal number of examples from each class (subjective and objective). For instance, in the 6-shot setting, the prompt included the three most similar (or dissimilar) subjective examples and the three most similar (or dissimilar) objective ones. This balance helps prevent prompt-induced bias toward a particular class label.

4.3. Multi-Agent LLM Reasoning

We design a set of multi-agent prompting experiments to investigate the interpretability and robustness of LLM outputs:

- **Debate setup:** Two agents argue why a sentence is subjective vs. objective; a third model acts as a judge.
- **Adversarial reasoning:** One agent argues why the sentence is not subjective and another why it is not objective, and a judge makes the final call.
- **Extended framing:** We include all four perspectives—Subjective, Not Subjective, Objective, and Not Objective— and a judge making the final decision.

5. Results

This section presents a detailed evaluation of multiple modeling strategies for subjectivity classification, including fine-tuned transformers, prompted LLMs with and without few-shot examples, advanced prompt reframing, and agent-based debating approaches. Given the dataset's imbalance, the official primary evaluation metric is the macro-averaged F1-score, which equally weights both classes. We also pay particular attention to SUBJ recall, as the subjective class is often underrepresented and may be more informative in downstream analyses. This imbalance was mitigated using a weighted BCE loss, warmup training, and early stopping to ensure stability and generalizability. We first report results with different system variants on the development subset and then discuss the official results obtained with the 2025 test set.

5.1. Preliminary Results

5.1.1. Fine-Tuned Transformers

Table 2 summarizes the performance of supervised transformer models trained on English and Arabic data. RoBERTa-Base, fine-tuned on English data, achieves the best performance overall, with a macro F1 of 0.70 and a notably high macro precision of 0.79. However, the model struggles with subjective instances, achieving only 0.39 recall for the subjective class, which suggests a strong bias toward the majority (objective) class.

In Arabic, the results are considerably weaker across the board. While BERT-Base-Arabertv02 achieves the best Arabic macro F1 score (0.55), subjective recall remains modest (0.47). Despite the use of language-specific models and XLM-RoBERTa for cross-lingual encoding, the performance gap between English and Arabic remains substantial.

Table 2
Fine-tuned transformer performance on subjectivity classification in English and Arabic.

Model	Setup	Lang	Macro F1	Macro P	P Subj	R Subj
RoBERTa-Base	10e, 5e-6 lr, 32 bs	English	**0.70**	0.79	0.76	0.39
XLM-RoBERTa	10e, 5e-6 lr, 32 bs	English	0.66	0.74	0.68	0.34
BERT-Base-Arabertv02	10e, 5e-5 lr, 16 bs	Arabic	**0.55**	0.55	0.50	0.47
XLM-RoBERTa	10e, 5e-6 lr, 16 bs	Arabic	0.51	0.54	0.51	0.25

5.1.2. Prompt Engineering and Few-Shot Learning

Table 3 presents results from zero- and few-shot prompting using GPT-4o-mini. A basic prompt yields poor subjective precision (0.32) but strong recall (0.67), suggesting that under-specification of the task leads the model to over-predict subjectivity. By contrast, a more detailed prompt derived from annotation guidelines improves both macro F1 (+0.12) and subjective precision (+0.14).

Incorporating six or twelve few-shot examples drawn randomly further boosts performance, reaching a macro F1 of 0.76. The improvement is consistent across both subjective precision and recall. Using 12-shot few-shot learning with extended prompting slightly improves recall but saturates macro F1 gains.

Table 3
Prompt-based and few-shot GPT-4o-mini performance in English.

System	Macro F1	Macro P	P Subj	R Subj
GPT-4o-mini (Basic Prompt)	0.54	0.57	0.32	0.67
GPT-4o-mini (Extended Prompt)	0.66	0.65	0.46	0.56
+ FSL (6-shot, Random)	**0.76**	0.78	0.69	0.60
+ FSL (12-shot, Random)	**0.76**	0.77	0.66	0.63

5.1.3. Few-Shot Selection Strategies

We further compare several strategies for selecting few-shot examples: random sampling, similarity-based sampling, and dissimilarity-based sampling. In similarity-based sampling, we choose examples that are most similar to the test sentence, while in dissimilarity-based sampling, we select those that are the most different. Similarity is measured using the cosine similarity between sentence embeddings generated by GPT-3. Results are shown in Table 4.

Interestingly, random selection outperforms similarity-based strategies across all models. For GPT-4o-mini, random sampling yields the best macro F1 (0.76), whereas dissimilarity-based selection offers a better recall (0.73) but slightly lower overall F1. A similar trend is observed for Qwen-72B, where

dissimilar sampling boosts recall (+0.07 over similarity) but offers minimal F1 gain. This suggests that dissimilar examples may help capture broader linguistic variance, aiding generalization. These results contrast with earlier findings highlighting the benefits of semantically similar examplars for in-context learning [32].

Table 4
Few-shot selection strategies across LLMs in English.

System	Macro F1	Macro P	P Subj	R Subj
GPT-4o-mini				
+ Random	**0.76**	0.78	0.69	0.60
+ Similarity	0.70	0.69	0.52	0.62
+ Dissimilarity	0.75	0.74	0.57	**0.73**
LLaMA 70B				
+ Random	0.73	0.73	0.61	0.57
+ Similarity	0.70	0.71	0.58	0.51
+ Dissimilarity	**0.75**	0.77	0.67	0.31
Qwen 72B				
+ Random	0.71	0.71	0.55	0.60
+ Similarity	0.71	0.70	0.52	0.67
+ Dissimilarity	**0.73**	0.72	0.57	0.64

5.1.4. Prompt Reframing and Debate-Based Inference

We investigate whether the way labels are framed affects model behavior. Reframing "subjective vs. objective" as a binary question (e.g., "Is the sentence subjective? Yes/No") or as category labels ("Category 1 vs. Category 2") leads to slight F1 gains over the base prompt (Table 5). Framing clearly influences the model's inductive bias, with category labels yielding better subjective precision (0.69) and macro F1 (0.72).

Debating-based prompting (Table 6) also provides strong results. The setup where one LLM argues for subjectivity, another for objectivity, and a judge decides, achieves the best macro F1 overall (0.77). Notably, this format significantly enhances subjective recall (up to 0.74), suggesting that reasoning-focused prompting facilitates more balanced decisions. Debate variants using negated prompts (e.g., "Not Subjective" vs. "Not Objective") also perform competitively.

Table 5
Effect of prompt reframing with GPT-4o-mini in English.

Framing Strategy	Macro F1	Macro P	P Subj	R Subj
Yes/No Binary	0.71	0.70	0.52	0.70
Category 1 vs 2	**0.72**	0.76	0.69	0.47

Table 6
Performance of debating-style LLMs with GPT-4o-mini in English.

Debating Setup	Macro F1	Macro P	P Subj	R Subj
Subjective vs Objective	**0.77**	0.76	0.62	0.72
Not Subjective vs Not Objective	0.76	0.75	0.59	**0.74**
Full Scale (Pos/NPos/Neg/NNeg)	0.74	0.73	0.56	0.74

5.1.5. LLM Ensemble Results

Finally, we evaluate an ensemble voting strategy that aggregates predictions from five diverse models: RoBERTa-Base, GPT-4o-mini, LLaMA 70B, Qwen 72B, and Aya-Expanse 32B. As shown in Table 7, this ensemble achieves the highest overall macro F1 score (0.79), with a strong subjective precision of 0.77. These results indicate that ensembling models with heterogeneous architectures and training paradigms can effectively capture complementary perspectives on subjectivity, enhancing robustness and performance.

Table 7
Ensemble classification performance in English.

System	Macro F1	Macro P	P Subj	R Subj
LLM Ensemble	**0.79**	0.77	0.77	0.59

5.2. Final Results

5.2.1. Evaluation Setup

The official evaluation of the CheckThat! 2025 campaign comprised three settings:

- **Monolingual**: train and test on data in a given language L (Arabic, Italian, German, English).
- **Multilingual**: train and test on data comprising several languages.
- **Zero-shot**: train on several languages and test on unseen languages (Romanian, Polish, Ukrainian, Greek).

For our final submitted system in the official campaign evaluation, we adopted the extended-prompt strategy using randomly selected 6-shot examples, paired with an ensemble of multiple models, including GPT-4 variants (GPT-4o-mini, GPT-4.1-mini), RoBERTa, LLaMA 70B, and Qwen 72B. In the zero-shot setting, the in-context examples were provided in English, following the task guidelines.

Final results are reported in Table 8.

5.2.2. Discussion

Our team demonstrated strong results across multiple languages, achieving first place in Arabic and Polish, and top-three positions in the majority of the evaluated languages, including Italian, English, and multilingual settings. This consistent performance underscores the robustness and generalizability of our approach using LLM-based few-shot learning.

Our experiments demonstrate that leveraging large language models (LLMs) instead of fine-tuned smaller language models (SLMs) can yield highly competitive results across multiple languages and settings. However, the effectiveness of LLMs critically depends on the quality of prompt design. For instance, advanced prompting strategies such as debating LLMs, where multiple model outputs are cross-examined, did not lead to substantial improvements over standard few-shot prompting. Similarly, varying the example selection method, whether by similarity, dissimilarity, or random choice, showed no significant impact on final performance. These findings suggest that while prompt engineering remains essential, more complex example selection or ensemble strategies may not always provide additional gains.

The most notable result was in Arabic, where we outperformed the second-ranked team by a substantial margin of +0.10 Macro F1-score. We attribute this advantage partly to the nature of the Arabic dataset, which exhibits annotation inconsistencies. Unlike fine-tuned models that heavily depend on high-quality labeled training data, our few-shot LLM approach is less affected by such noise. Prior research has indicated that in-context learning with LLMs can be relatively independent of the exact label quality provided in training examples [33]. Consequently, our method was more resilient to

Table 8
Final Macro F1 results for each language including the top-ranked team, **CEA-LIST** (our team), the baseline, and the last-ranked team. When **CEA-LIST** is first, the second-ranked team is also reported.

Language	Team	Rank	Macro F1
Italian	XplaiNLP	1	0.8104
	CEA-LIST	**2**	**0.8075**
	Baseline	11	0.6941
	IIIT Surat	14	0.4612
Arabic	**CEA-LIST**	**1**	**0.6884**
	UmuTeam	2	0.5903
	Baseline	8	0.5133
	JU_NLP	14	0.4328
German	smollab	1	0.8520
	CEA-LIST	**4**	**0.7733**
	Baseline	15	0.6960
	IIIT Surat	16	0.6342
English	msmadi	1	0.8052
	CEA-LIST	**3**	**0.7739**
	UGPLN	22	0.5531
	Baseline	23	0.5370
Multilingual	TIFIN India	1	0.7550
	CEA-LIST	**3**	**0.7396**
	Baseline	13	0.6390
	AI Wizards	16	0.2380
Polish	**CEA-LIST**	**1**	**0.6922**
	IIIT Surat	2	0.6676
	Baseline	9	0.5719
	TIFIN INDIA	14	0.3811
Ukrainian	CSECU-Learners	1	0.6424
	Baseline	5	0.6296
	CEA-LIST	**10**	**0.6061**
	TIFIN INDIA	14	0.4731
Romanian	msmadi	1	0.8126
	CEA-LIST	**6**	**0.7659**
	Baseline	13	0.6461
	TIFIN INDIA	14	0.5181
Greek	AI Wizards	1	0.5067
	CEA-LIST	**7**	**0.4492**
	Baseline	9	0.4159
	TIFIN India	14	0.3337

inconsistencies, resulting in superior evaluation performance. This highlights a significant practical benefit of using LLMs: they can better handle noisy or imperfect datasets, offering an edge in real-world scenarios where high-quality annotations are difficult to obtain.

Overall, our findings suggest that LLMs, combined with carefully crafted few-shot prompts, offer a powerful and flexible alternative to traditional fine-tuning approaches, especially when training data quality varies. This has important implications for future multilingual sentiment analysis tasks and other NLP challenges where data quality and multilingual coverage are key concerns.

6. Dataset quality

During our evaluation of the Arabic dataset, we consistently observed limited performance across all tested configurations. Regardless of the prompting strategy employed, ranging from simple to extended prompts, in both zero-shot and few-shot settings, the macro F1-score remained below **0.55**. This plateau in performance was observed across multiple models, including GPT-4 and fine-tuned BERT models as

shown in Table 2, and suggests potential issues beyond model capacity or prompt design.

Interestingly, this contrasts with the results reported in the original dataset paper [13], where a five-shot setup using a Maximal Marginal Relevance (MMR)-based example selection achieved an F1-score of **0.80**. In our comparable few-shot setting with GPT-4 and similarity-based example selection, the F1-score reached only **0.547**, indicating a significant gap in reproducibility.

Upon closer inspection, we identified potential sources of annotation inconsistency. Manual review revealed several instances where the assigned labels did not seem to align with the guidelines outlined in the original paper. For example, the sentence:

"مع كلّ عزيمة وعنفوان أبطال مخيم جنين وأهل جنين وأهل فلسطين يجب أن تتكامل الساحات الأخرى معهم، وأن يعرف العدوّ ومن يؤيّده أنّ الشعب الفلسطيني ليس وحده."

"With all the determination and fervor of the heroes of Jenin camp, the people of Jenin, and the people of Palestine, other fronts must unite with them, and the enemy and those who support it must know that the Palestinian people are not alone."

is labeled as objective, despite the presence of emotive language and the term "the enemy" (ٱلْعَدُوُّ), which could reasonably be interpreted as subjective under the dataset's own criteria, where they state that such politically charged language must be labeled as subjective. Conversely, clearly factual sentences such as:

شاهدوا الآن البث المباشر لافتتاح كأس العالم 2024عبر لبنان 24

"Watch now Lebanon 24 live broadcast via 'Lebanon 24' for the opening of the 2022 World Cup."

are labeled as subjective, despite appearing to report straightforward event announcements. Similarly, sentences comprising purely reported speech, such as the following about COVID-19 statistics, are labeled as subjective even though the annotation guidelines specify otherwise:

"زيادة وفيات وإصابات كورونا في إيران خلال الـ24 ساعة الماضية بالتزامن مع إعلان رئيس مركز الأمراض المعدية بوزارة الصحة عن تراجع تدريجي في إحصاءات كورونا، أعلنت إدارة العلاقات العامة بالوزارة عن زيادة أخرى في عدد الوفيات والإصابات خلال الـ24 ساعة الماضية."

"An increase in COVID-19 deaths and infections in Iran over the past 24 hours, coinciding with the announcement by the head of the Infectious Diseases Center at the Ministry of Health of a gradual decline in COVID-19 statistics, while the Public Relations Department of the Ministry announced another increase in the number of deaths and infections during the past 24 hours."

To ensure this was not a limitation inherent to the task or language, we ran comparable experiments on the Arabic dataset from the 2023 edition of the task. In that case, our models achieved significantly better performance (F1 = 0.84) using a six-shot extended prompt, and the top team of that year had reported an F1-score of 0.79 [34], demonstrating the feasibility of high performance on well-annotated Arabic datasets.

To further test the hypothesis that label quality rather than linguistic features was the bottleneck, we translated the dataset into English using DeepL and reran the experiments. However, this also did not lead to improved performance (F1 < 0.6), reinforcing our initial hypothesis. A small-scale manual reannotation conducted by one of the authors, who is a native Arabic speaker, led to a moderate increase in performance (F1 = 0.65), providing further evidence that inconsistencies in labeling may play a role in the observed results.

These observations highlight the challenges of subjectivity annotation, especially in politically sensitive contexts, and underline the importance of annotation consistency for benchmarking tasks involving subtle linguistic distinctions.

7. Conclusion

In this study, we have shown that large language models (LLMs) used with well-designed few-shot prompting can rival or surpass fine-tuned smaller models (SLMs) across diverse languages and settings. Our approach proved particularly robust in the face of noisy or inconsistent training data, as demonstrated by our strong performance on the Arabic dataset. By consistently ranking among the top teams, securing first place in Arabic and Polish and top-three finishes in most other languages, we illustrate the versatility and effectiveness of LLMs for multilingual subjectivity detection.

While advanced prompt engineering strategies such as debating and varied example selection did not yield major improvements, our results emphasize the critical role of prompt quality itself. The flexibility of LLMs combined with minimal reliance on extensive labeled data offers a promising path forward for multilingual NLP tasks, especially when dealing with data of varying quality.

Acknowledgments

This work was supported by the BOOM ANR Project (ANR-20-CE23-0024) and benefited from the use of the FactoryIA supercomputer, funded by the Île-de-France Regional Council.

Declaration on Generative AI

This work was assisted by generative AI tools used to improve clarity and style, specifically GPT-4. The authors reviewed and verified all content to ensure accuracy and maintain the integrity of the scientific work.

References

[1] F. Ruggeri, F. Antici, A. Galassi, A. Korre, A. Muti, A. Barron, et al., On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection, in: CEUR Workshop Proceedings, volume 3370, CEUR-WS, 2023, pp. 103–111.

[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] B. Abimbola, E. de La Cal Marin, Q. Tan, Enhancing legal sentiment analysis: A convolutional neural network–long short-term memory document-level model, Machine Learning and Knowledge Extraction 6 (2024) 877–897. URL: https://www.mdpi.com/2504-4990/6/2/41. doi:10.3390/make6020041.

[4] B. Liu, Sentiment analysis and opinion mining, Springer Nature, 2022.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[6] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. M. Struß, F. Antici, M. Hasanain, J. Köhler, et al., Overview of the clef-2023 checkthat! lab: Task 2 on subjectivity in news articles, in: 24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023, CEUR Workshop Proceedings (CEUR-WS. org), 2023, pp. 236–249.

[7] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, et al., Overview of the clef-2024 checkthat! lab task 2 on subjectivity in news articles, in: CEUR Workshop Proceedings, volume 3740, CEUR-WS, 2024, pp. 287–298.

[8] 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, Faggioli et al.[22] (2024).

[9] S. Javdan, B. Minaei-Bidgoli, et al., Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection, in: Proceedings of the second workshop on figurative language processing, 2020, pp. 67–71.
[10] W. Zhang, Y. Deng, B. Liu, S. Pan, L. Bing, Sentiment analysis in the era of large language models: A reality check, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3881–3906. URL: https://aclanthology.org/2024.findings-naacl.246/. doi:10.18653/v1/2024.findings-naacl.246.
[11] 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.
[12] T. Huang, E. Fan, Structured reasoning for fairness: A multi-agent approach to bias detection in textual data, 2025. URL: https://arxiv.org/abs/2503.00355. arXiv:2503.00355.
[13] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, Thatiar: subjectivity detection in arabic news sentences, arXiv preprint arXiv:2406.05559 (2024).
[14] Kusrini, M. Mashuri, Sentiment analysis in twitter using lexicon based and polarity multiplication, in: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 365–368. doi:10.1109/ICAIIT.2019.8834477.
[15] S. Zahoor, R. Rohilla, Twitter sentiment analysis using lexical or rule based approach: A case study, in: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 537–542. doi:10.1109/ICRITO48877.2020.9197910.
[16] A. Kotelnikova, D. Paschenko, K. Bochenina, E. Kotelnikov, Lexicon-based methods vs. bert for text sentiment analysis, in: International Conference on Analysis of Images, Social Networks and Texts, Springer, 2021, pp. 71–83.
[17] R. A. Potamias, G. Siolas, A.-G. Stafylopatis, A transformer-based approach to irony and sarcasm detection, Neural Computing and Applications 32 (2020) 17309–17320.
[18] G. Marvin, N. Hellen, D. Jjingo, J. Nakatumba-Nabende, Prompt engineering in large language models, in: International conference on data intelligence and cognitive informatics, Springer, 2023, pp. 387–402.
[19] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, arXiv preprint arXiv:2402.07927 (2024).
[20] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, K. Chen, Prosa: Assessing and understanding the prompt sensitivity of llms, 2024. URL: https://arxiv.org/abs/2410.12405. arXiv:2410.12405.
[21] F. Errica, G. Siracusano, D. Sanvito, R. Bifulco, What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering, 2025. URL: https://arxiv.org/abs/2406.12334. arXiv:2406.12334.
[22] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.
[23] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Forty-first International Conference on Machine Learning, 2024. URL: https://openreview.net/forum?id=zj7YuTE4t8.
[24] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, Z. Tu, Encouraging divergent thinking in large language models through multi-agent debate, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17889–17904. URL: https://aclanthology.org/2024.emnlp-main.992/. doi:10.18653/v1/2024.emnlp-main.992.
[25] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, Advances in Neural Information Processing Systems 36 (2023) 8634–8652.
[26] 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.
[27] 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. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
[28] P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, P. Nakov, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, arXiv preprint arXiv:1808.05542 (2018).
[29] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness., CLEF (Working Notes) 2380 (2019).
[30] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media., CLEF (working notes) 2696 (2020).
[31] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, et al., Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates., in: CLEF (working notes), 2021, pp. 369–392.
[32] B. Xu, Q. Wang, Z. Mao, Y. Lyu, Q. She, Y. Zhang, *k*nn prompting: Beyond-context learning with calibration-free nearest neighbor inference, 2023. URL: https://arxiv.org/abs/2303.13824. arXiv:2303.13824.
[33] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work?, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 11048–11064. URL: https://aclanthology.org/2022.emnlp-main.759/. doi:10.18653/v1/2022.emnlp-main.759.
[34] 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: CLEF (Working Notes), 2023, pp. 318–328.

A. Prompts Used

In this section, we report the prompts used for the classification of the sentences. We report both the simple prompt we experimented with at first and the extended prompt that provided the best performance. We also report the prompts used for the debating LLMs. In particular, we report the prompt used for : (1) the LLM tasked with explaining why a sentence is objective ,(2) the LLM tasked with explaining why a sentence is subjective ,(3) the LLM tasked with explaining why a sentence is not objective, (4) the LLM tasked with explaining why a sentence is not subjective and (5) the judge LLM.

A.1. Simple Prompt (English):

You are a linguistic expert, able to detect whether a sentence is objective (OBJ) or subjective (SUBJ). Answer only with `OBJ` or `SUBJ`.

A.2. Extended Prompt (English):

You are a linguistic expert specializing in detecting whether a sentence is objective or subjective. Your task is to classify sentences according to the following criteria:

- **Objective:** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker. I repeat: **SENTENCES WHICH ONLY CONTAIN REPORTED SPEECH SHOULD NEVER BE LABELED SUBJECTIVE.**
- **Subjective:** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., 'so damaged') can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like 'will hope to sow uncertainty' suggest an interpretation rather than a fact.

Answer only with the words `objective` or `subjective` based on these criteria.
Note: For other languages, this extended prompt was translated using DeepL to ensure semantic accuracy and consistency.

A.3. Subjectivity Explanation Prompt:

You are a linguistic expert specializing in detecting whether a sentence is objective (OBJ) or subjective (SUBJ). Your task is to classify sentences according to the following criteria:

- **Objective:** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker. I repeat: **SENTENCES WHICH ONLY CONTAIN REPORTED SPEECH SHOULD NEVER BE LABELED SUBJECTIVE.**
- **Subjective:** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., "so damaged") can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like "will hope to sow uncertainty" suggest an interpretation rather than a fact.

Given the following sentence, explain why it is classified as **subjective** based on these criteria. Try to be concise and explain why it is classified as such. Do not repeat the sentence in your answer. Keep to the annotation guidelines given above. Maintain a critical mindset, you can disagree with the classification, but do so only if you are certain. Do not speculate about the sentence's intention.

A.4. Objectivity Explanation Prompt:

You are a linguistic expert specializing in detecting whether a sentence is objective (OBJ) or subjective (SUBJ). Your task is to classify sentences according to the following criteria:

- **Objective:** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker. I repeat: **SENTENCES WHICH ONLY CONTAIN REPORTED SPEECH SHOULD NEVER BE LABELED SUBJECTIVE.**
- **Subjective:** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., "so damaged") can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like "will hope to sow uncertainty" suggest an interpretation rather than a fact.

Given the following sentence, explain why it is classified as **objective** based on these criteria. Try to be concise and explain why it is classified as such. Do not repeat the sentence in your answer. Keep to the annotation guidelines given above. Maintain a critical mindset, you can disagree with the classification, but do so only if you are certain. Do not speculate about the sentence's intention.

A.5. Non-Subjectivity Explanation Prompt:

You are a linguistic expert specializing in detecting whether a sentence is objective (OBJ) or subjective (SUBJ). Your task is to classify sentences according to the following criteria:

- **Objective:** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker. **SENTENCES WHICH ONLY CONTAIN REPORTED SPEECH SHOULD NEVER BE LABELED SUBJECTIVE.**
- **Subjective:** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., "so damaged") can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like "will hope to sow uncertainty" suggest an interpretation rather than a fact.

Given the following sentence, explain why it should **not** be classified as **subjective** based on these criteria. Try to be concise and explain why it does not fit the criteria for subjectivity. Do not repeat the sentence in your answer. Focus only on why it fails to meet the conditions for subjectivity.

A.6. Non-Objectivity Explanation Prompt:

You are a linguistic expert specializing in detecting whether a sentence is objective (OBJ) or subjective (SUBJ). Your task is to classify sentences according to the following criteria:

- **Objective:** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker. **SENTENCES WHICH ONLY CONTAIN REPORTED SPEECH SHOULD NEVER BE LABELED SUBJECTIVE.**
- **Subjective:** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., "so damaged") can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like "will hope to sow uncertainty" suggest an interpretation rather than a fact.

Given the following sentence, explain why it should **not** be classified as **objective** based on these criteria. Try to be concise and explain why it does not fit the criteria for subjectivity. Do not repeat the sentence in your answer. Focus only on why it fails to meet the conditions for objectivity.

A.7. Judge Prompt:

You are a judge LLM tasked with determining whether a sentence is objective (OBJ) or subjective (SUBJ) based on opinions defending different points of view. Your job is to evaluate these opinions according to the following criteria:

- **Objective (OBJ):** A sentence is objective if it presents factual information, even if the information is debatable or controversial. Additionally:
  - **Emotions:** Statements conveying emotions should be labeled as objective if they reflect the author's beliefs or sensations that cannot be fact-checked or rephrased in a more neutral form.
  - **Quotes:** If a sentence contains a direct quote, label it as objective, since the task concerns only the subjectivity of the article's author, not the quoted speaker.
- **Subjective (SUBJ):** A sentence is subjective if it reflects personal opinions, interpretations, or evaluations. Indicators of subjectivity include:
  - **Intensifiers:** Words or phrases that amplify a statement (e.g., *so damaged*) can indicate subjectivity, as they may reflect the author's personal perspective.
  - **Speculations:** Statements that imply uncertainty, predictions, or unverifiable claims should be labeled as subjective. For example, phrases like *will hope to sow uncertainty* suggest an interpretation rather than a fact.
- **Edge Cases:**
  - **Emotions:** Although statements carrying emotions convey a subjective point of view, they cannot be verified or confuted by a fact-checking system and are therefore labeled as objective.

- **Quotes:** When authors use quotes to support their thesis, the quoted content may be subjective, but for classification, it should be labeled as objective, focusing only on the article's author.
- **Intensifiers:** The presence of intensifiers can indicate subjectivity, but it's important to assess whether they genuinely reflect the author's perspective or serve a descriptive purpose.
- **Speculations:** Speculative statements should be regarded as subjective, as they often reflect the author's interpretation and not just factual content.

Given the sentence and the opinions, your task is to make a final decision and answer only with `objective` or `subjective`.