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

Akram Elbouanani, *Evan Dufraisse, Aboubacar Tuo, Adrian Popescu*

akram.elbouanani@cea.fr, adrian.popescu@cea.fr

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Motivation

- **Why subjectivity detection matters: Essential for fact-checking, media analysis, moderation: distinguishing opinion from fact is critical.**
- **The LLM advantage:** Traditional SLMs require extensive annotated data; LLMs with prompting may offer greater flexibility and robustness when data is scarce or noisy.
- **How can we optimize LLM performance?**

Results

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

Language	Team	Rank	Macro F1
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

Baseline

- We fine-tune a simple RoBERTA model and use it as a baseline for comparison.

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

Prompting Strategies

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.

LLMs tended to struggle with this one…

Prompting Strategies

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

- The extended prompt improves performance. Adding few-shot examples improves it even further.
- Performance seems to plateau at 6 shots.

Prompting Strategies

- Can we go further than that through a better selection of the few-shot examples?
- We test three strategies:
  - Randomly selecting few-shot examples.
  - Selecting the most similar train sentences to the current test sentence.
  - Selecting the most dissimilar train sentences to the current test sentence.

Prompting Strategies

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

- There aren’t any big notable differences.
- A very interesting result is that the quality of labels does not seem to impact performance!

Prompting Strategies

- What if we reframe the labels?

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

- Reframing the labels does not improve performance in English.
- However, translating labels or using numerals for labels improves performance for certain other languages.

Debating LLMs

- Debating LLM Systems is an emerging paradigm to enhance LLM performance.

- We try out different settings for our debates:
  - One LLM arguing **for** a “subjective” answer, one LLM arguing **for** an “objective” answer.
  - One LLM arguing **against** a “subjective” answer, one LLM arguing **against** an “objective” answer.
  - We include all four perspectives: “subjective”, “not subjective”, “objective”, and “not objective”.
  - A judge LLM makes the final call.

Debating LLMs

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

- Debating LLMs only seem to marginally change performance.

Ensemble

- What if we just ensemble a bunch of models?

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

- Just throw a bunch of LLMs at it!

Discussion

- LLMs outperform SLMs on subjective detection, especially with few-shot prompting.
- Arabic dataset: noisy annotations hurt SLMs; LLMs handled it better and won by a clear margin.
- Takeaway: LLMs are robust and adaptable, even on messy data, though more resource-heavy.

Thank you !