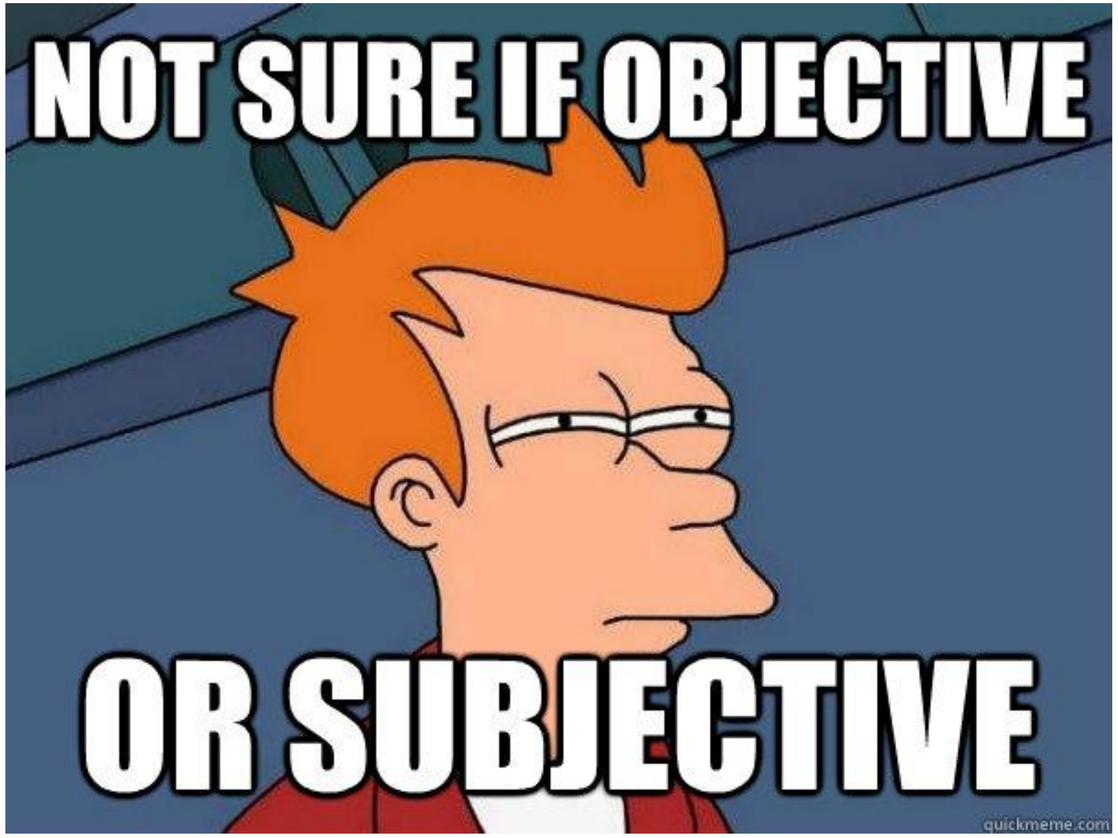
XplaiNLP at CheckThat! 2025: Multilingual Subjectivity Detection with Finetuned Transformers and Prompt-Based Inference with Large Language Models

Authors: Ariana Sahitaj, Jiaao Li, Pia Wenzel Neves, Fedor Splitt, Premtim Sahitaj, Charlott Jakob, Veronika Solopova, Vera Schmitt

Problem Statement

- Subjective language can hide in factual reporting making it hard to detect especially across languages with different signals and limited resources.



[1]

Problem Statement

- Subjective language can hide in factual reporting making it hard to detect especially across languages with different signals and limited resources.

- **Problem / Task:**
  How do we distinguish between subjective and objective language across languages?

Objective or Subjective?

"The proposed agreement includes the best wage increases for rail workers in over forty years."

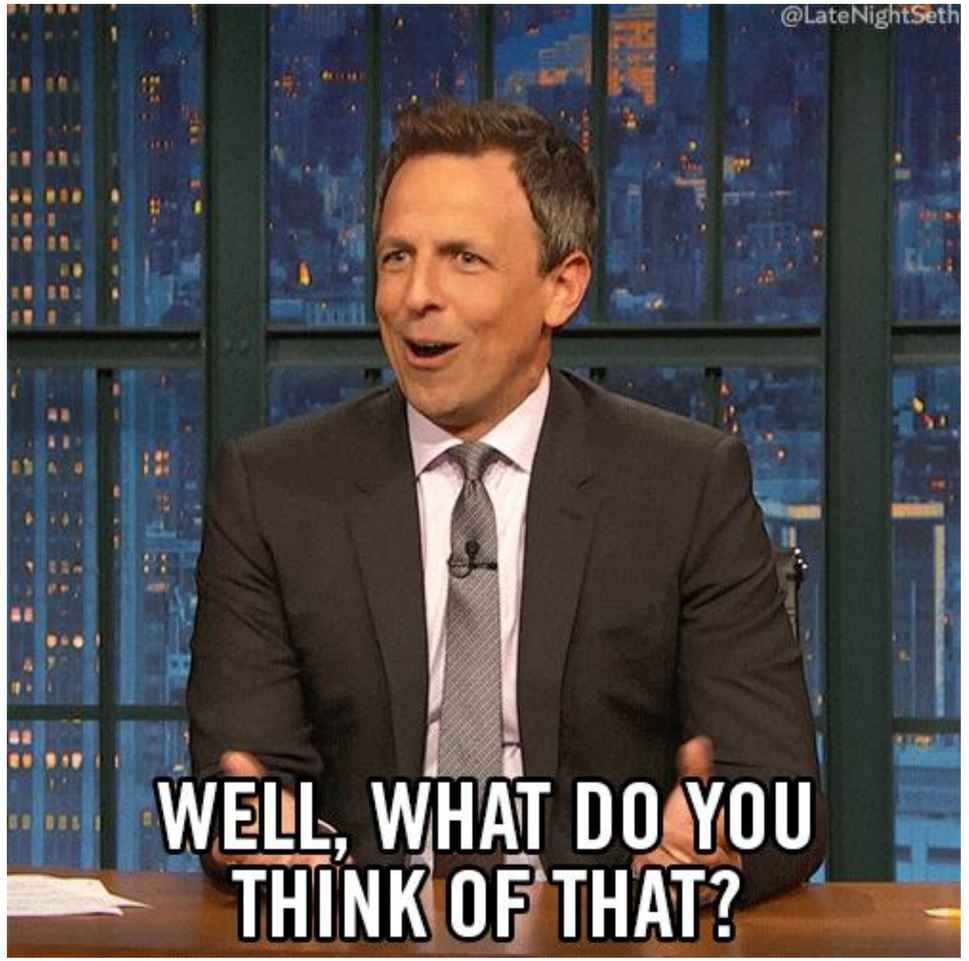
id: d6b59bd5-dda4-4adc-bf5f-51aa1ee91d4e

OBJ SUBJ

"Was she really advocating a nuclear first strike?"

id: c06d89e9-bf9d-42bb-9ee9-ec62f0cb14ed

OBJ SUBJ



[2]

Dataset

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

English training data is ~64% objective, Italian even stronger bias toward objectivity (>73% OBJ)

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

Arabic and Greek (zero-shot) have the longest average sequences, while English and Italian are shorter

Ambiguities and Hard-to-Translate Cases


[3]

- **Manual analysis** of English sentences with annotation conflicts to identify common sources of labeling disagreement.
  - **Immigration Discourse**
    - *“Mr. Buchanan’s criticism of immigration”*
  - **Race and Social Commentary**
    - *“CRT anti-white curricula”* or *“diversity, equity, and inclusion”*
  - **Media and Political Rhetoric**
    - *“ Lügenpresse”* or *“Treason Lobby”*
  - **Sarcasm and Pragmatic Devices**
    - *“What could possibly go wrong?”*
  - **Framing in Economic and Environmental Topics**
    - *“ bribed by a globalist billionaire”*

Fine-Tuned Transformers

Fine-tuning German-BERT with Translated Training Data

[4]

- Expanded **German** training set by **translating other languages** (English, Italian, Bulgarian, Arabic) and adding them in order of linguistic similarity.
- **Incremental data addition** improved performance up to Bulgarian; adding Arabic decreased F1-score.
- Best results achieved with **German-BERT fine-tuned on German, English, Italian, and Bulgarian** data.

Training Configuration and Performance Across Languages

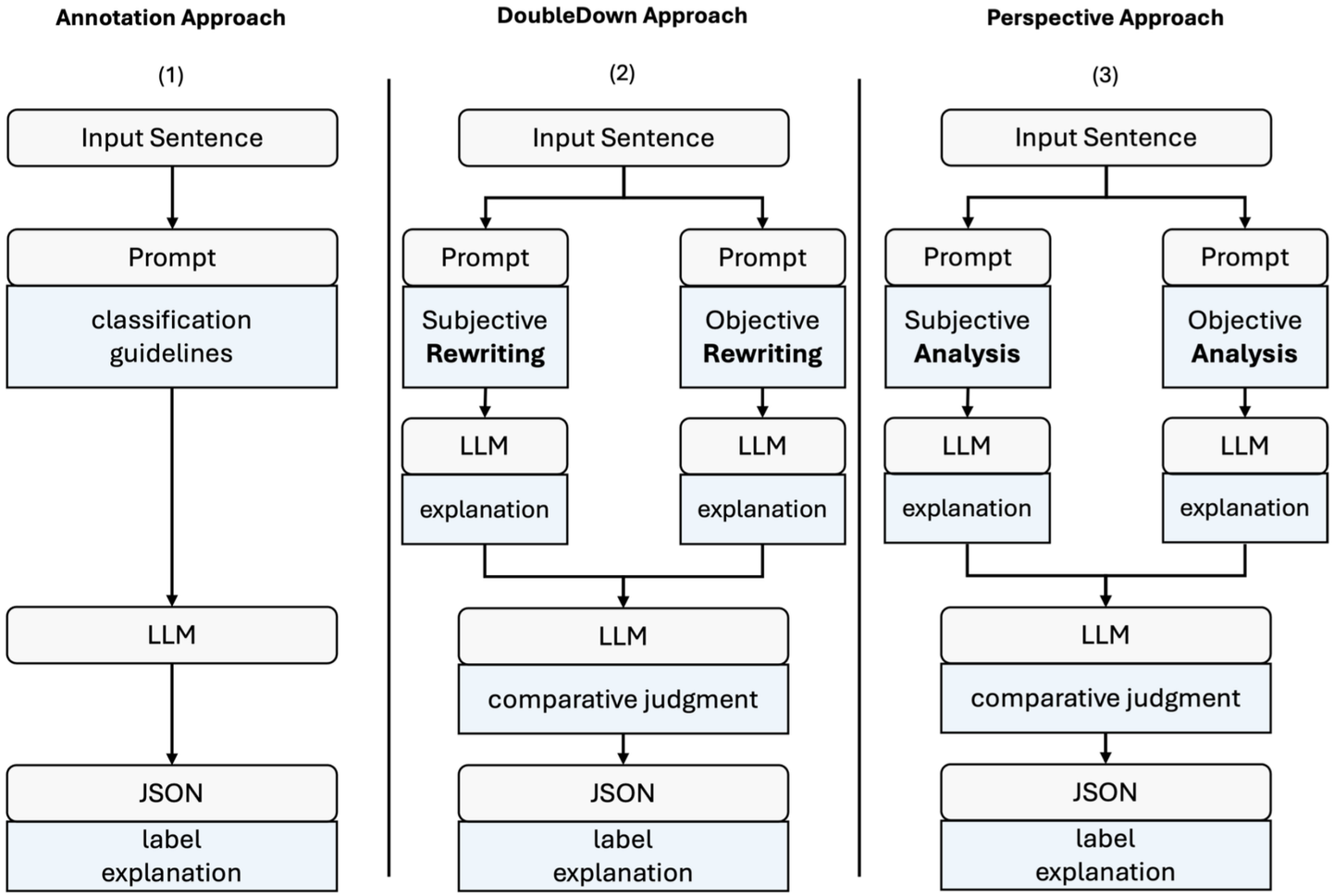
Languages	Sentences	Epochs	Learning Rate	DEV macro-F1	DEV-TEST macro-F1
de	800	4	0.001	0.7253	-
de, en	1630	4	0.001	0.7405	-
de, en, it	3243	5	0.001	**0.7651**	0.7712
de, en, it, bg	3972	5	0.001	**0.7692**	**0.8172**
de, en, it, bg, ar	6418	5	0.004	0.7275	-

Monolingual Fine-tuning EuroBERT and XLM-RoBERTa-base

- Fine-tuned **EuroBERT** and **XLM-RoBERTa-base** for monolingual subjectivity detection, using datasets in all languages except Arabic.
- Applied **class imbalance handling** with Focal Loss, class weighting, and reduced threshold (0.45) for SUBJ prediction.
- Training setup included **early stopping**, temperature scaling, mixed-precision training, and gradient clipping for stability and calibrated predictions.

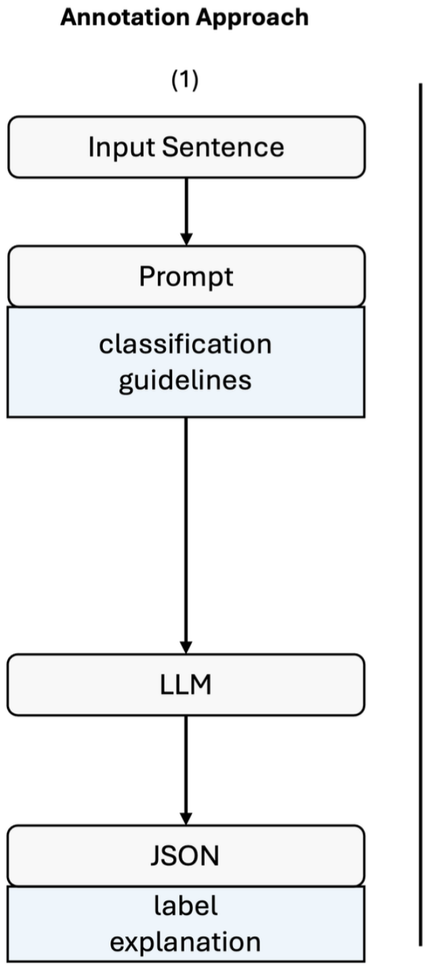
[5]

LLM-based Inference Strategies



- (1) the LLM classifies the input sentence **based on explicit guidelines** provided in the prompt
- (2) **rewrites** the **sentence in** both **subjective and objective styles** and infers the label by comparing which version better preserves the original meaning
- (3) **generates explanations for both possible interpretations** and selects the label based on the more convincing reasoning.

LLM-based Inference Strategies



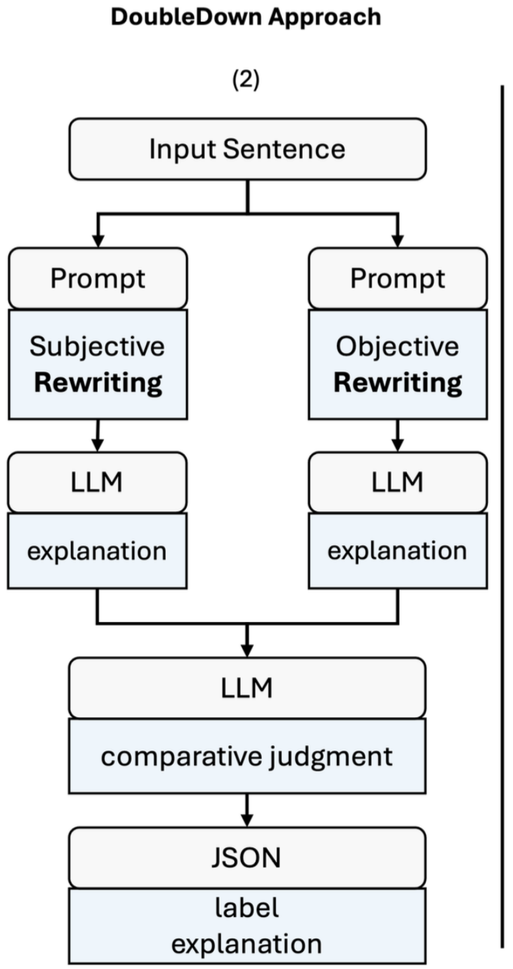
Input Sentence: "Blanco established himself earlier in his career working for Dr. Luke's Kasz Money Productions."

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

- Model provided with **14 decision rules** defining linguistic and rhetorical cues for subjectivity vs. objectivity.
- Extracted **verdict** and **supporting explanation** from the model's output.

o3-mini

LLM-based Inference Strategies



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

Reframes classification as a two-way comparison instead of a direct label

- **subjective** (opinion, personal evaluation or emotional language)
- **objective** (factual content, neutral phrasing)

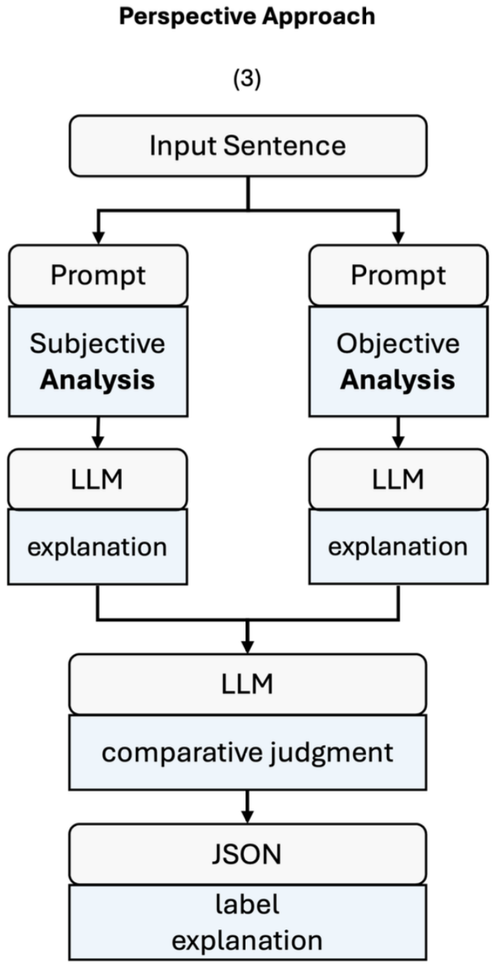
- input sentence + both rewrites → version best preserves the original meaning

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.

Label: OBJ

gpt-4.1-mini

LLM-based Inference Strategies



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

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

gpt-4.1-mini

Results

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

- Rows are sorted by how close our F1 score is to the top result.
- Bold values in our F1 indicate scores above the baseline.

Evaluation results for our approach across subtasks, compared to the baselines and top-ranked scores

Two complementary approaches proved effective:

- **Prompting** excelled in high-resource languages (e.g., Italian → 1st place, English > baseline).
- **Fine-tuned Transformers (esp. XLM-RoBERTa)** delivered the most consistent results in multilingual and zero-shot tasks.

Important limitations shaped the outcomes:

- Only **one submission per language** was allowed → model choice was pragmatic rather than optimal.

Implications, Limitations and Future Work

Implications:

- Both fine-tuned transformer models and zero-shot prompting with LLMs can be effective for subjectivity detection
- Performance depending on language and resource availability

Limitations:

- Submission rules restricted testing of all model–approach combinations
- No broader context given
- Label imbalance across languages risked bias and over-prediction of subjectivity

Future Directions:

- Explore hybrid strategies (ensembling, few-shot prompting, confidence-based switching) to improve performance in low-resource and zero-shot settings.

Questions?

CONTACT US
FOR FURTHER QUESTIONS

Ariana Sahitaj
Student Researcher
ariana.sahitaj@campus.tu-berlin.de

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,

 Image Sources

[1]http://www.quickmeme.com/img/6b/6be2f3dbe2c99178c5a2a58c9a02105c87662e22dc65b042834eaa2e34c83343.jpg
[2]https://media3.giphy.com/media/v1.Y2lkPTc5MGI3NjExbXdrY3QzdXkwajB2emE1eWpuenphd2U2NzNremg3ZXJqYjFqdnpkZyZlcD12MV9pbnRlcm5hbF9naWZfYnlfaWQmY3Q9Zw/fimQBzrS2rCL73lxxt/giphy.gif
[3] https://www.complex.com/sneakers/a/robert-facey/buyer-beware-the-20-craziest-fake-kicks-on-the-internet_220511
[4] https://huggingface.co/google-bert/bert-base-german-cased
[5] https://www.linkedin.com/posts/matthias-kindt-486498232_eurobert-ist-ein-multilinguales-offenes-activity-7305105225776812033-FURa?originalSubdomain=de