nlu@utn at CheckThat! 2025: Combining Bias Sensitivity, Linguistic Features, and Persuasion Cues in an Ensemble for Subjectivity Detection

Selina Meyer[1,*], Michael Roth[1]

[1]University of Technology Nuremberg, Dr.-Luise-Herzberg-Straße 4, 90461 Nürnberg, Germany

Abstract

This paper details our participation in task 1 on subjectivity detection at 2025's CheckThat! lab. Specifically, we focused on monolingual English data provided by the task organizers. Our approach consisted of an ensemble model, made up of a) a classifier pretrained on political bias data and fine-tuned on the subjectivity training data, b) a BERT classifier fine-tuned on linguistically augmented training data, and c) a BERT classifier fine-tuned on training data augmented with persuasion techniques used in the text. The final label was determined using majority voting. Our approach resulted in a Macro F1 score of 74.86% on the test set, ranking 7^{th} place on the task leaderboard. Based on additional experiments conducted over the course of our participation, our ensemble-based system outperformed state-of-the-art large language models, including Google's Gemini-Flash and GPT-4o-Mini on this task, highlighting that, given the presented experimental setup, this type of task is still a challenge for generative AI.

Keywords

subjectivity detection, credibility, natural language processing

1. Introduction

Subjective writing in news can be a marker of partisan and biased reporting but can also be a stylistic and narrative device to increase the readability of articles [1]. udging subjectivity is no easy task—even for human experts—as it is shaped by subjective interpretation. Nevertheless, Antici et al. [2] have introduced a corpus of sentences from news articles judged as objective and subjective, garnering high inter-annotator agreement through a multistep approach. Their corpus consists of articles sourced from British news outlets with diverse political orientations and annotators were tasked to annotate on a sentence-to-sentence basis, without being given context for the respective sentences. In our participation in the English subtask of Task 1 [3] presented at CheckThat! 2025 [4, 5], we draw on the insights from related research on news credibility to improve classifier performance.

According to Antici et al.'s definition used for annotation, a "sentence is considered subjective when it is based on—or influenced by—personal feelings, tastes, or opinions" [2]. This betrays a close relation to the concept of credibility, which is often defined by research-based credibility cues in the context of natural language processing [6]. Such cues can be context-based, but also content-based, meaning identifiable on the basis of the text itself. Content-based credibility signals include partisanship and bias, linguistic cues, general quality characteristics of text and clickbaitiness of headlines, presence of logical fallacies, and use of persuasion techniques [6]. Earlier work, especially before the introduction of transformer-based models, has also heavily focused on the use of lexical and other linguistic information [6].

Other submissions to this task in previous years have used linguistic features, and most have used transformer models as underlying architectures [7]. Casanova et al. [8], who ranked first place in 2024's edition of this task, used an ensemble approach based on two transformer models, as well as subjectivity scores based on an independently developed expert system for subjectivity and vagueness detection.

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

*Corresponding author.

selina.meyer@utn.de (S. Meyer); michael.roth@utn.de (M. Roth)

0000-0002-4736-2565 (S. Meyer); 0000-0002-9128-519X (M. Roth)

In our approach to the task, we build on the previously seen success of using ensemble models for subjectivity detection. In contrast to Casanova et al.'s [8] approach, each of the three models in our ensemble is informed by a different aspect of cues which are likely to inform subjectivity detection classifiers. Specifically, we leverage a model trained on the related task of persuasion technique detection [9] to augment the training data before fine-tuning a BERT-instance, fine-tune a BERT-based classifier previously fine-tuned on a bias detection task [10, 11], and combine this with a third BERT-instance fine-tuned on augmented data based on our own linguistic analysis of the provided training data. Our ensemble model reaches a Macro F1 score of 74.86% on the test set, ranking 7^{th} place on the official task leaderboard, with only 0.58% separating us from 5^{th} place. This is an improvement of 21 percentage points over the baseline provided by the task organizers.[1]

2. Methods

Below, we describe the two data augmentation approaches we employed for this task, as well as the classifiers and fine-tuning involved in creating the ensemble model.

2.1. Linguistic Analysis and Augmentation

Gajewska [12] explored the role of pronouns and modal verbs, as well as emotional and polarising words in subjectivity detection, comparing the predictive power of naive models trained with these features with BERT-based methods. In their experiments, simple BERT-finetuning on the training data, with no additional meta-features, outperformed all other approaches tested in their experiments, resulting in a Macro F1 score of 0.70 on the test-set. This indicates, that a naive approach to linguistic feature selection is not enough to successfully inform subjectivity classification. Instead, we analyze the training data to identify linguistic features which statistically differentiate subjective from objective sentences.

To achieve this, we tokenize the training data using spaCy and parse the simple and detailed part-of-speech (POS) tags, syntactic dependencies, and named entity labels for each token [13]. Following this, we identify significant differences in the frequency of each of these features between the subjective and objective class. Table 1 shows all features with significantly differing frequency between the two classes, and their corresponding bonferroni-corrected p-values.

Table 1
SpaCy features with significant differences in frequency between subjective and objective class, based on analysis of simple POS, detailed POS, dependency tags and entity labels.

Feature	p-val	Feature	p-val
Adjective (ADJ)	0.028	Adverb (ADV)	0.001
Auxiliary (AUX)	0.005	Numeral (NUM)	<0.001
Pronoun (PRON)	<0.001	Proper Noun (PROPN)	<0.001
Cardinal Number (CD)	<0.001	adjective (English), other noun-modifier (Chinese) (JJ)	0.02
Verb. modal auxiliary (MD)	0.002	Noun, proper singular (NNP)	<0.001
pronoun, personal (PRP)	0.004	Adverb (RB)	<0.001
Verb, base form (VB)	<0.001	Verb, past tense (VBD)	<0.001
adverbial modifier (advmod)	0.001	Compound (compund)	0.001
Numeric modifier (nummod)	0.005	modifier of quantifier (quantmod)	0.004
Absolute or relative dates or periods (DATE)	<0.001	Companies, agencies, institutions, etc. (ORG)	<0.001
Percentages, including "%" (PERCENT)	0.004		

Based on these insights, the original sentences in the training data were augmented with a [SEP] token after the original sentence, followed by the numerical count of each differentiating feature in the sentence (see Table 2).

[1]Code available at https://github.com/SelinaMeyer/nlu-utn_at_CheckThat2025

Table 2
Example for training data augmented with linguistic features and persuasion techniques.

Linguistically-informed augmentation
Gone are the days when they led the world in recession-busting[SEP]ADJ-count: 0; ADV-count: 0; AUX-count: 1; NUM-count: 0; PRON-count: 1; PROPN-count: 0; CD-count: 0; JJ-count: 0; MD-count: 0; NNP-count: 0; PRP-count: 1; RB-count: 0; VB-count: 0; VBD-count: 1; advmod-count: 1; compound-count: 1; nummod-count: 0; quantmod-count: 0; DATE-count: 0; ORG-count: 0; PERCENT-count: 0; '
Persuasion technique-based augmentation
Gone are the days when they led the world in recession-busting[SEP]['Loaded Language'][SEP][[0, 62]][SEP]['0.83782977']

2.2. Persuasion Technique-Based Augmentation

Subjectivity in writing can manifest in the use of loaded language, name calling, appeal to values, use of faulty argumentation, e.g. strawman or red herring arguments, and other types of biased writing. These techniques have been explored by other researchers in the context of persuasion and propaganda detection [14, 15]. To explore whether knowledge of such techniques can enhance the performance of subjectivity classifiers, we used a model created in the context of a shared task in SemEval 2023 [9] available via API access on GATE Cloud[2]. The model is passed a piece of text and returns the persuasion techniques detected in the sentence, along with confidence scores and the start and end character of the span for which a respective persusasion technique was identified. We augment the training data with the returned information by appending it with a list of identified persuasion techniques, followed by the list of spans and confidence scores, each separated by [SEP] tokens (See Table 2).

2.3. Model Training

Our ensemble method employs three models: $\text{BERT}^{\text{PERS}}$ is fine-tuned on the persuasion-augmented training data, $\text{BERT}^{\text{LING}}$ is fine-tuned on the linguistically-augmented training data. Both of these are based on instances of `bert-base-uncased` [16][3]. In contrast, $\text{BERT}^{\text{BIAS}}$ is based on a pre-trained BERT model available on Hugging Face, as provided by Baly et al. [10][4], which was initially fine-tuned on news articles labeled with their displayed political ideology (right, center, left). For the subjectivity detection task, we further fine-tuned this model using the non-augmented training data. Since the original classifier was designed for a multi-class task with three categories, we adapted the classification head to accommodate binary classification. After this modification, we froze the weights of the initial layers and only fine-tuned the final four layers. In the resulting ensemble model, the class of a text sample is decided via simple majority voting between the three models.

We use the same hyperparameter configuration for $\text{BERT}^{\text{LING}}$ and $\text{BERT}^{\text{PERS}}$ during fine-tuning, applying a learning rate of 2e-5, weight decay of 0.01, a batch size of 16 and training over 20 epochs. To fine-tune $\text{BERT}^{\text{BIAS}}$ we reduce the learning rate to 1e-5. An early stopping callback was implemented for all models with a patience of three epochs. The best model, based on performance on the dev-split was loaded at the end of training. Fine-tuning was performed on a NVIDIA A40 GPU and took between 74.68 and 125.83 seconds per model. The early stopping callback resulted in a training of 11 epochs for $\text{BERT}^{\text{BIAS}}$, 10 epochs for $\text{BERT}^{\text{PERS}}$, and 6 epochs for $\text{BERT}^{\text{LING}}$.

3. Results

Table 3 displays the classification performance of the individual fine-tuned models, the ensemble model, and the baseline, a logistic regression model trained on Sentence-BERT multilingual embeddings provided by the task organizers, on the provided dev, dev-test and test splits.

[2] https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier
[3] https://huggingface.co/google-bert/bert-base-uncased
[4] https://huggingface.co/bucketresearch/politicalBiasBERT

Table 3
Model performance of individual classifiers and ensemble model on the Dev, Dev-Test and Test Split as calculated by the scorer provided by task organizers and performance displayed on the task leaderboard.

Model	Dev Split	Dev-Test Split	Test Split
$\text{BERT}^{\text{PERS}}$+$\text{BERT}^{\text{BIAS}}$+$\text{BERT}^{\text{LING}}$	**0.80**	**0.71**	**0.75**
$\text{BERT}^{\text{PERS}}$	0.77	0.69	0.72
$\text{BERT}^{\text{BIAS}}$	0.73	0.68	0.69
$\text{BERT}^{\text{LING}}$	0.76	**0.71**	0.69
Baseline	0.73	0.63	0.54

The ensemble approach led to a Macro F1 increase of 7% over the baseline on the dev split, 8% on the dev-test split, and 21% on the test split as returned by the scorer provided by the task organizers. Performance on the test set as displayed on the official leaderboard was at Macro F1 of 74.86% placing our submission in 7^{th} place.

4. Ablation Studies

We conduct a series of ablation studies to assess the necessity of the individual classifiers in the ensemble, compare the performance of our ensemble model with state-of-the-art large language models (LLMs)—namely Google Gemini, Google Gemma, and GPT-4—and explore an LLM-in-the-loop approach to confirm or reject the predictions made by the ensemble.

4.1. Replacing Individual Classifiers

To test the role the individual classifiers in the ensemble play in classification performance, we replace each classifier in turn with a bert-base-uncased instance, fine-tuned on the non-augmented training data ($\text{BERT}^{\text{BASELINE}}$). The results are presented in Table 4.

Table 4
Macro F1 Scores of Different Ensemble Variations on Dev and Test Splits.

Ensemble-version	Dev Split	Dev-Test Split	Test Split
$\text{BERT}^{\text{PERS}}$+$\text{BERT}^{\text{LING}}$+$\text{BERT}^{\text{BASELINE}}$	0.77	**0.72**	0.72
$\text{BERT}^{\text{BIAS}}$+$\text{BERT}^{\text{LING}}$+$\text{BERT}^{\text{BASELINE}}$	**0.80**	0.71	**0.73**
$\text{BERT}^{\text{PERS}}$+$\text{BERT}^{\text{BIAS}}$+$\text{BERT}^{\text{BASELINE}}$	0.78	0.70	0.72
$\text{BERT}^{\text{BASELINE}}$	0.76	0.70	0.71

Interestingly, although $\text{BERT}^{\text{BIAS}}$ alone leads to the lowest classification results of all the classifiers included in our ensemble model, replacing it with $\text{BERT}^{\text{BASELINE}}$ yields the largest discrepancy in classification results compared to the original ensemble on the dev split, whereas removing $\text{BERT}^{\text{PERS}}$, which on its own yields the highest score on the dev split, does not decrease the classification score on the dev split at all compared to the ensemble results and leads to the lowest decrease in Macro F1 on the test split. On the dev-test split, replacing $\text{BERT}^{\text{BIAS}}$ with $\text{BERT}^{\text{BASELINE}}$ leads to a Macro F1 increase of 1%, whereas replacing $\text{BERT}^{\text{LING}}$ decreases the score by the same amount and on the test split, replacing any of the three classifiers in the ensemble leads to a decrease in classification scores, with the difference being smallest when $\text{Bert}^{\text{PERS}}$ is replaced. To explore potential reasons for this, we calculate the agreement rate between the individual classifiers on the test set, presented in Table 5.

The classifications produced by $\text{BERT}^{\text{PERS}}$ largely align with those of $\text{BERT}^{\text{BASELINE}}$, which accounts for the minimal change in results when the two models are substituted for one another. At the same time, $\text{BERT}^{\text{LING}}$ and $\text{BERT}^{\text{BIAS}}$ show larger discrepancies with the remaining models, suggesting they contribute more variability to the predictions. This added diversity may explain why replacing them with $\text{BERT}^{\text{BASELINE}}$ results in larger discrepancies in classification scores for the dev and test splits.

Table 5
Agreement Rate Between Individual Classifiers in Ensemble Model and $BERT^{BASE}$

	$BERT^{BASELINE}$	$BERT^{BIAS}$	$BERT^{PERS}$	$BERT^{LING}$
$BERT^{BASELINE}$	1	80%	91.33%	77.33%
$BERT^{BIAS}$		1	78%	72.67%
$BERT^{PERS}$			1	73.33%
$BERT^{LING}$				1

4.2. Generative LLM-based Classification

Given that the data at hand is hard to annotate even for humans [2], we were interested in how generative LLMs perform at this task. We conducted a range of experiments employing two proprietary LLMs, namely Google's `gemini-2.0-flash-lite` (Gemini Flash) and OpenAI's `gpt-4o-mini-2024-07-18` (GPT-4o-mini), as well as the smaller, open-source LLM `gemma2-9b-it` (Gemma2-9b), also created by Google, which can be run on a single GPU. We test different prompt conditions, inspired by the annotation guidelines and examples used for the training data provided by Antici et al. [2]. To save resources, the full range of experiments was conducted only with Gemini-Flash on the dev split. The prompt yielding the best classification results utilizing Gemini-Flash was then also run using Gemma2-9b to explore whether smaller, locally run models can compare to large scale model performance on this task, and GPT-4o-mini to identify whether using a different model with comparable size can boost performance. Finally, the best performing system was used to predict on the dev-test and test splits.

The prompts provided for the models were each made up of a system prompt, telling the model that it is an expert at judging the subjectivity of texts, followed by different prompt variants explained below, and finally an instruction to provide the response in JSON format, containing a label, an explanation for why the label was given, the text fragment that was most important for the decision, and, depending on the prompt variant, the subjectivity or objectivity subclass.

The following prompt variants were explored:

- *Simple prompt* - The model was instructed to classify sentences as subjective or objective, with no detailed definitions of what constitutes subjectivity or objectivity, to identify the model's inherent ability to differentiate between subjective and objective texts.
- *Full guidelines* - The full guidelines provided in Antici et al. [2], including the definitions of different subclasses of subjectivity and objectivity, were adapted to the use of LLMs and passed to the model.
- *Few-shot-style* - The examples for various objectivity and subjectivity subclasses provided in the guidelines by Antici et al. [2] were restructured into a few-shot learning format and passed to the model.
- *Guidelines without examples* - All examples provided in the guidelines were removed, and only the definitions of each subjective and objective subclass were passed to the model.

The exact wording for each of the conditions is provided in Appendix A. The results of this range of experiments are presented in Table 6.

In the *Full guidelines* condition, Gemini-Flash achieved classification performance on the dev split, which had been used as a validation set during fine-tuning of the BERT-based models, that closely approached that of the ensemble model. This initially suggested that the LLM-based system might outperform the ensemble classifier on the dev-test and test splits. However, this expectation was not met. On the dev-test split, Gemini-Flash's Macro F1 score was 6 percentage points lower than that of the ensemble, suggesting that the dev-test split is generally more challenging to classify than the remaining splits. This assumption is further supported by the consistently higher results yielded by tested models on the test split relative to the dev-test split. On the test split, Gemini-Flash's classification score was 3 percentage points lower than the ensemble's result.

Table 6
LLM Performance for Different Prompting Conditions.

Model	Condition	Split	Macro F1
Gemini-Flash	Simple prompt	Dev	0.66
	Full guidelines	Dev	**0.78**
	Few-shot-style	Dev	0.73
	guidelines without examples	Dev	0.75
GPT-4-o-mini	Full guidelines	Dev	0.73
Gemma2-9b	Full guidelines	Dev	0.70
Gemini-Flash	Full guidelines	Dev-Test	0.65
Gemini-Flash	Full guidelines	Test	0.72

4.3. Generative LLM-in-the-Loop

Finally, we test whether generative LLMs could be used to improve ensemble-based classifications. To this end, we pass Gemini-Flash the annotation guidelines (as in the *full guidelines* condition above) along with a sample classified by the ensemble model and information about the model's performance and the label distribution found in the data. The model is then prompted to confirm or reject the ensemble model's classification, based on the annotation guidelines. Since an error analysis of the LLM experiments above revealed that a frequent error source was the model's inability to differentiate between the author's opinion and that of a quoted third-party, a sentence was added to the instructions to increase awareness of this. This approach led to a decrease in classification scores compared to using only the ensemble model across the dev, dev test and test splits, leading to Macro F1 scores of 0.75, 0.68, and 0.72, respectively.

5. Error Analysis and Discussion

As mentioned above, our error analysis of generative LLM classifications showed that Gemini-Flash struggled to differentiate between the opinions of the author and those of third parties quoted in the text when prompted using the annotation guidelines. An analysis of misclassified samples of the ensemble model in the test split revealed a bias towards the subjective class. 68.18% of misclassifications were false positives.

Based on a closer analysis of the misclassified samples in the test set, it appears that many of the sentences constitute observations about a prominent person or event, as well as personal narratives either based on the author's experiences or not discernible as quotations in the sentence itself (see examples in Table 7). Although not directly subjective, these samples appear to exhibit a narrative style, which could make them harder to classify.

Table 7
Examples for Misclassified Sentences in the Test Split

Sentence	Label	Prediction
His actors create their characters initially based on people they know, then flesh them out through a long process of development, rehearsal and improvisation — a method pioneered on the clock of the 1970s BBC that today makes his films unusually hard to finance.	OBJ	SUBJ
Despite the hardships, however, revolutionary optimism was palpable.	OBJ	SUBJ
The bing-bong when someone presses the "stop" button on the city bus, and the accompanying sing-songy announcements in Korean, the beeps of riders scanning their transit cards to board or depart; soju-drunk office workers loudly singing off-tune through neighbourhood alleyways; obnoxiously loud K-pop music blaring out of storefronts; and songs that seem to change key at record rates as delivery motorbikes speed out of range.	OBJ	SUBJ
I would soothe my loneliness and isolation in the evenings by playing endless hours of Law & Order: Special Victims Unit just for the ambient sound – the comfort of Detectives Olivia Benson and Elliot Stabler bringing criminals of the worst kind to justice.	OBJ	SUBJ

Although our approach did not achieve top performance compared to other submissions, we draw a couple of interesting conclusions from our experiments and ablation studies, which we discuss

here. The first is that, although generative LLMs have led to many advancements in natural language understanding, subjectivity detection given the current experimental setup still poses a challenge even to top-performing LLMs, such as GPT-4 and Google Gemini. The task requires a lot of nuance and is hard to annotate even for humans [2], which appears to have limited LLM performance in our experiments. We also find indications that Gemini-Flash's internal representation of subjectivity may differ significantly from the guidelines developed by Antici et al. [2]. This is reflected in the remarkably low performance of the *simple prompt* condition in our LLM experiments, where the model judged subjectivity based solely on its own understanding. Future work could investigate this discrepancy more systematically, for example by explicitly probing the model's internal criteria for subjectivity across diverse contexts and comparing different LLMs.

Furthermore, our ensemble method achieved better classification scores independently than in combination with the generative LLM-in-the-loop. This suggests that for certain classification tasks and experimental setups, in which generative LLMs do not perform particularly well on their own, incorporating them in this manner does not lead to improvements. A possible direction for future work is testing out the capabilities of LLMs in this context more extensively, for instance through dynamic selection of few-shot samples, as presented for example in An et al. [17] or other state-of-the-art prompting approaches, which might significantly increase LLM-performance for this task.

We also find that explicitly providing linguistic information relevant to a specific dataset, as we did in the training of the $\text{BERT}^{\text{LING}}$ classifier, can be beneficial for classification performance, even though transformer-based models already implicitly encode such information [18], underlined by the results of the replacement ablation study, which showed that $\text{BERT}^{\text{LING}}$ appeared to be the most impactful classifier in the ensemble. The positive impact of leveraging linguistic features, as operationalized by POS-tags and similar automatically retrievable information, for judging subjectivity has been shown by Nakashole and Mitchell [19] in the past, although not in combination with transformer-based approaches. Future work could explore whether supplying this type of information to generative LLMs might improve classification results.

6. Conclusion

In this paper, we presented our approach to the English subtask of Task 1 of CheckThat!2025. Our ensemble approach used three BERT-based models, each informed by linguistic differences between the subjective and objective classes in the training data, research on persuasion techniques in propaganda, and bias in news articles. The system placed in the upper third of the English leaderboard.

Our ablation studies showed that using linguistic information in an explicit way contributed to the ensemble and that models, which perform poorly by themselves can still contribute to an ensemble approach, as was the case for $\text{BERT}^{\text{BIAS}}$ in our experiments. Experiments with generative LLM-based classification and LLM-in-the-loop methods showed that the task remains a challenge for generative AI, given the presented experimental setup.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Paraphrase and reword, improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Steensen, Subjectivity as a journalistic ideal, in: B. K. Fonn, H. Hornmoen, N. Hyde-Clarke, Y. B. Hågvar (Eds.), Putting a Face on it: Individual Exposure and Subjectivity in Journalism, Cappelen Damm Akademisk, 2017, pp. 25–47.

[2] 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/.

[3] 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, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[4] 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.

[5] 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.

[6] I. Srba, O. Razuvayevskaya, J. A. Leite, R. Moro, I. B. Schlicht, S. Tonelli, F. M. García, S. B. Lottmann, D. Teyssou, V. Porcellini, et al., A survey on automatic credibility assessment of textual credibility signals in the era of large language models, arXiv preprint arXiv:2410.21360 (2024).

[7] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, et al., Overview of the clef-2024 checkthat! lab: check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 28–52.

[8] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier, P. Égré, Hybrinfox at checkthat! 2024-task 2: Enriching bert models with the expert system vago for subjectivity detection, in: CLEF 2024-Conference and Labs of the Evaluation Forum, 2024, pp. 1–9.

[9] B. Wu, O. Razuvayevskaya, F. Heppell, J. A. Leite, C. Scarton, K. Bontcheva, X. Song, SheffieldVeraAI at SemEval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1995–2008. URL: https://aclanthology.org/2023.semeval-1.275/. doi:10.18653/v1/2023.semeval-1.275.

[10] R. Baly, G. Da San Martino, J. Glass, P. Nakov, We can detect your bias: Predicting the political ideology of news articles, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), EMNLP '20, 2020, pp. 4982–4991.

[11] Political bias classification using finetuned bert model (2023).

[12] E. Gajewska, Eevvgg at checkthat! 2024: evaluative terms, pronouns and modal verbs as markers of subjectivity in text, Faggioli et al.[22] (2024).

[13] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020).

[14] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news articles, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: https://aclanthology.org/D19-1565/. doi:10.18653/v1/D19-1565.

[15] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup,

in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2343–2361. URL: https://aclanthology.org/2023.semeval-1.317/. doi:10.18653/v1/2023.semeval-1.317.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[17] S. An, B. Zhou, Z. Lin, Q. Fu, B. Chen, N. Zheng, W. Chen, J.-G. Lou, Skill-based few-shot selection for in-context learning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 13472–13492. URL: https://aclanthology.org/2023.emnlp-main.831/. doi:10.18653/v1/2023.emnlp-main.831.

[18] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3651–3657. URL: https://aclanthology.org/P19-1356/. doi:10.18653/v1/P19-1356.

[19] N. Nakashole, T. Mitchell, Language-aware truth assessment of fact candidates, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 1009–1019.

A. Prompts

A.1. System Prompt

```
You are an expert at judging the subjectivity of texts and are skilled at
adhering to guidelines provided to you.
For the text provided to you, first judge whether it is objective or subjective.
Then compare your initial judgement to each of the provided objectivity and
subjectivity criteria to determine whether your initial judgement is correct.
If your initial judgement is correct, return the label you have chosen.
If your initial judgement is incorrect, return the label that is correct
according to the guidelines.
```

A.2. Simple Prompt

```
You are tasked with judging whether a sentence is subjective or objective.
Give your answer in a JSON dictionary with the following format:
"explanation": [a short explanation of why the sentence is subjective or
objective]
"decisive text fragment": [the text snippet that is the most important for
deciding between objective and subjective]
"label": [SUBJ for subjective or OBJ for objective]
Judge the following sentence:
{text}
```

A.3. Full Guidelines

```
You are tasked with judging whether a short text is subjective or objective.
You are handed the guidelines below, given in triple backticks, to inform
```

your decision.
Guidelines:
```
{guidelines (Appendix A.7)}
```

Give your response in the following JSON-Format:
"label": [SUBJ for subjective or OBJ for objective],
"explanation": [a short explanation, why the sentence is objective or subjective],
"decisive text fragment": [the text snippet that is the most important for your decision],
"case": [The specific case of objectivity or subjectivity],

Judge the following sentence according to the guidelines provided above:
{text}

A.4. Guidelines without Examples

You are tasked with judging whether a short text is subjective or objective. You are handed the guidelines below, given in triple backticks, to inform your decision.

Guidelines:
```
Definitions
**Subjective.** A sentence is considered subjective when it is based on —or influenced by— personal feelings, tastes, or opinions. Otherwise, the sentence is considered objective.

**Specific Subjective Cases**
SUBJ 1.
A sentence is subjective if it explicitly reports the personal opinion of its author. Rhetorical questions are considered as an expression of an opinion as well; see Ex. (c). Additionally, speculations which draw conclusions are considered as opinions.

SUBJ 2.
A sentence is subjective if it contains sarcastic or ironic expressions attributable to its author.
.... (see Appendix A.7 for remaining categories)
```

Give your response in the following JSON-Format:
"label": [SUBJ for subjective or OBJ for objective]
"explanation": [a short explanation, why the sentence is objective or subjective],
"decisive text fragment": [the text snippet that is the most important for your decision],
"case": [The specific case of objectivity or subjectivity],

Judge the following sentence according to the guidelines provided above:

{text}

A.5. Few-Shot Style

You are tasked with judging whether a sentence is subjective or objective.
Guidelines:
```
Definitions
**Subjective.** A sentence is considered subjective when it is based on —or influenced by— personal feelings, tastes, or opinions. Otherwise, the sentence is considered objective.
**Objective.** If a sentence does not meet any subjectivity type listed in the previous section, it is considered objective. Here we include examples of objective sentences which may be wrongly interpreted as subjective.

Text: India, who was the bridesmaid at the King's wedding to Princess Diana in 1981, could not be seen in the footage, but filmed the video as she walked through the grounds of the royal residence.
Label: Obj
Explanation: The sentence is objective because the author describes a historical event without giving any opinion or personal comment

Text: It is a sad truth that many of the villages in this region of Portugal are dying.
Label: Subj
Explanation: The sentence is subjective because the author explicitly conveys their personal emotions, making the sentence subjective.

...(see Appendix A.7 for remaining categories)
```

Answer in the following JSON-Format:
"explanation": [a short explanation of why the sentence is subjective or objective],
"decisive text fragment": [the text snippet that is the most important for deciding between objective and subjective],
"label": [SUBJ for subjective or OBJ for objective],

Judge the following sentence:
{text}

A.6. LLM-as-a-Judge

You are tasked with judging whether the subjectivity prediction of a short text given by a classifier with 80% accuracy is correct or not.
The data contains around twice as many objective sentences as subjective sentences.
You are handed the guidelines below, given in triple backticks, to inform your decision.

Guidelines:

“‘
{guidelines}
“‘

Be mindful of the presence of quotation marks and other markers that indicate the presence of a third-party opinion.
Such markers should be considered as explicit third-party opinions and make the sentence objective, even if they are not clearly stated in the sentence.
The texts given to you for judgement are labelled as SUBJ for subjective and OBJ for objective.
Given the guidelines above, judge the correctness of the subjectivity label of the following Text:
Text: {text}
Label: [label]

Give your response in the following JSON-Format:
"explanation": [a short explanation of why the label is correct or incorrect, referring to the guidelines],
"correct": [True if the label is correct, False if it is incorrect]

A.7. Guidelines, based on [2]

Definitions
Subjective. A sentence is considered subjective when it is based on —or influenced by— personal feelings, tastes, or opinions. Otherwise, the sentence is considered objective.
The following Sentence is objective because the author describes a historical event without giving any opinion or personal comment:
Ïndia, who was the bridesmaid at the King’s wedding to Princess Diana in 1981, could not be seen in the footage, but filmed the video as she walked through the grounds of the royal residence.¨
In contrast, in the next Sentence the author explicitly conveys their personal emotions, making the sentence subjective:
Ït is a sad truth that many of the villages in this region of Portugal are dying.¨

Specific Subjective Cases
SUBJ 1.
A sentence is subjective if it explicitly reports the personal opinion of its author. Rhetorical questions are considered as an expression of an opinion as well; see Ex. (c). Additionally, speculations which draw conclusions are considered as opinions, see Ex. (d).
Examples:
(a) It has everything you could want in a holiday: beautiful sandy beaches and clear waters, ancient history and culture, delicious food (the Greek salads are simply on another level), island hopping, nightlife and more.
(b) After treading vineyard soils and seeing grapes ripening, that merlot becomes more than just a Wednesday night relaxant.
(c) Do they really think other nations sprouted up out of the ground?
(d) But Putin will hope to sow uncertainty in the eyes of policymakers’ meetings in New York.

SUBJ 2.
A sentence is subjective if it contains sarcastic or ironic expressions attributable to its author.
Examples:
(e) It’s no lie that the USA is one heck of a big country (said in a southern twang).
(f) With Land Rover bowdlerising images of the vehicle into little more than a perfume advertisement on TV[...].
(g) Especially if you’re more excited at the prospect of sampling rare bottles from the cellar than snapping vineyard selfies.

SUBJ 3.
A sentence is subjective if it contains exhortations of personal auspices made by its author.
Examples:
(h) The West should arm Ukraine faster.

SUBJ 4.
A sentence is subjective if it contains discriminating or downgrading expressions.
Examples:
(i) And what is even more evident is the perverse role reversal that is taking place, in which he who sits in Rome has the task of formulating heterodox principles opposed to Catholic doctrine, and his accomplices in the Dioceses have the role of scandalously applying them, in an infernal attempt to undermine the Moral law in order to obey the spirit of the world.
(j) How did we reach the stage where priests and bishops cowered like frightened puppies before a common flu, where their predecessors ministered fearlessly among the lepers, the cripples, and the victims of typhoid, cholera, smallpox, and Bubonic Plague?

SUBJ 5.
A sentence is subjective if it contains rhetorical figures, like hyperboles, explicitly made by its author to convey their opinion.
Examples:
(k) Barcelona where it all began, Messi was a king in Catalonia and he lived like one too.
(l) The churches, and the Catholic Church in particular (which is by far the largest), had the ability to put an end to the lockdown madness and the COVIDterror campaign, had they wished to do so.
(m) So it must be biochemistry that is really what is racist.

Specific Cases of Objectivity
If a sentence does not meet any subjectivity type listed in the previous section, it is considered objective. Here we include examples of objective sentences which may be wrongly interpreted as subjective.

Case 1.
A sentence is objective when it reports on news or historical facts that are quoted by the author of the sentence.
Examples:
(a) President Putin has just reiterated his threat to use nuclear weapons and

announced that Russiancontrolled Ukrainian territory will become part of the Russian Federation. (b) In the modern era electroconvulsive therapy, first used in 1938, became a treatment for some serious forms of depression in the post-war decades.

Case 2.

A sentence is objective when it describes the personal feelings, emotions or moods of the writer, without conveying opinions on other matters.

Examples:

(c) I was definitely surprised at how emotional I felt watching the service.

(d) The second I saw him, I felt a jolt of connection.

Case 3.

A sentence is objective if it expresses an opinion, claim, emotion or a point of view that is explicitly attributable to a third-party (e.g., a person mentioned in the text).

Examples:

(e) Frank Drake believed that the universe had to contain other intelligent beings.

(f) Ÿou showed callous indifference to Dean's fate after he had been repeatedly stabbedẗhe judge said.

Note: The presence of quotation marks (“ ”), when used to quote a third person (be it at the beginning
of the sentence, at the end, or both), represents an explicit third-party opinion, even if it is not clearly stated in the sentence.

Examples:

(g) “Crosbie is an extremely violent man who has no place in society, and we welcome the jury's verdict today.”

(h) “My children have lost their hero and I have lost my chosen person - the person I chose to spend my life with.

(i) For these reasons and out of conviction, I consider myself bound in my conscience to say no.”

Case 4.

A sentence is objective if it contains a comment made by the author of the sentence that does not draw any conclusion.

In particular, the author doesn't convey their personal interpretation or opinion, leaving the discussion on the topics of interest open.

Examples:

(j) It is not clear yet which of the couples from the E4 reality show remain together and who have now, because the series has not concluded.

(k) Do car manufacturers know how far their EVs will really go?

(l) Exact figures are hard to come by, but Ukraine may well have more troops available than Russia now.

Case 5.

A sentence is objective if it contains factual conclusions made by the author of the sentence that do not convey any stance or personal opinion, or are justified up by a non-personal hypothesis.

Examples:

(m) In years gone by, travel to Japan was notoriously expensive, but the devaluing of the yen has made it more accessible.

(n) The bottom-up approaches which target the molecular, genetic and

electrical fundamentals of the brain can assist top-down approaches to brain disorder such as talking therapies.
(o) Based on our experiences and road tests, a good rule of thumb is to expect to achieve somewhere between 75 and 80 per cent of a car's WLTP Combined range[. . .]

Case 6.
When referring to an individual, any kind of well-known nickname or title is considered objective.
Examples:
(p) Things have certainly progressed on the pitch for Spurs this season.
(q) The Duke of York 'plotted' with Diana to 'push Prince Charles aside'.

Case 7.
Any kind of common expression or proverb is considered objective.
Examples:
(r) the adage 'sticks and stones may break my bones, but words can never hurt me'.
(s) Home sweet home: George poses in one of the rooms at his sprawling Hampstead home during a photoshoot in 2002.