JU_NLP at CheckThat! 2025: A Confidence-guided Transformer-based Approach for Multilingual Subjectivity Classification

Notebook for the CheckThat! Lab at CLEF 2025

Srijani Debnath[1,*], Dipankar Das[2]

[1]Department of Computer Science & Engineering, Government College of Engineering and Leather Technology, Kolkata, India
[2]Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

Abstract

With the rapid progress of transformer-based multilingual models, accurately distinguishing subjective authorial viewpoints from objective reportage in news articles has become increasingly challenging. This paper presents a comprehensive framework for binary subjectivity classification that combines fine-tuned multilingual BERT with a lightweight, feature-based post-processing module for English. We fine-tuned `bert-base-multilingual-cased` on five languages—Arabic, Bulgarian, English, German, and Italian—under monolingual, multilingual, and zero-shot conditions, and evaluated on held-out data in Greek, Polish, Ukrainian, and Romanian. For the English monolingual subtask, low-confidence predictions (confidence<0.7) were refined using two lexical features: a SentiWordNet–based subjective score and a combined lexical-clues plus opinion-lexicon count. Our approach achieved macro-averaged F1 scores of 0.545 for Arabic (rank-14), 0.733 for English (rank-8), 0.699 for Italian (rank-10), 0.736 for German (rank-9), 0.435 for Greek (rank-8), 0.560 for Polish (rank-11), 0.580 for Ukrainian (rank-11), 0.744 for Romanian (rank-8), and 0.654 in the multilingual setting (rank-10).

Keywords

subjectivity classification, multilingual BERT, post-processing, macro F1

1. Introduction

Subjectivity detection in news articles has become an essential component of modern natural language processing pipelines. The ability to distinguish whether a sentence expresses an author's personal viewpoint or reports factual information underpins tasks such as sentiment analysis, bias detection, and information extraction. Transformer-based models like BERT[1] and XLM-R[2] have achieved impressive results on many cross-lingual benchmarks, yet classification of subjectivity remains challenging due to subtle linguistic cues, domain variation, and limited annotated resources in many languages.

The main contributions of this paper can be summarized as follows:

- We developed a subjectivity classification framework using transformer-based mBERT [3] model to classify under monolingual, multilingual, and zero-shot conditions across nine subtasks.
- We extracted two lexical features from English sentences: a subjective score computed via SentiWordNet synsets [4] and a combined count of lexical-clues and opinion-lexicon words [5].
- To further enhance English predictions, we introduced a lightweight post-processing module that overrides low-confidence outputs (confidence<0.7) using two interpretable lexical features: a SentiWordNet–based subjective score and a combined lexical-clues plus opinion-lexicon count.
- We demonstrated that our framework yields robust performance, achieving macro-F1 scores ranging from 0.435 to 0.744 across all monolingual and zero-shot subtasks and 0.654 in the multilingual setting.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
srijanidebnath2005@gmail.com (S. Debnath); dipankar.dipnil2005@gmail.com (D. Das)

2. Related Work

Early approaches to subjectivity and sentiment classification relied on shallow machine learning models with hand-crafted features. Pang and Lee (2004) employed n-gram features with Support Vector Machines to achieve strong performance on English sentiment data [6]. Lexicon-based methods, such as those using subjectivity lexicons by Wiebe et al. (2002), extracted polarity scores and worked well in monolingual settings but struggled to generalize across domains and languages [7].

With the advent of neural networks, convolutional and recurrent architectures became popular. Kim (2014) demonstrated that a simple CNN over word embeddings could outperform traditional feature-based models on various text classification tasks [8]. BiLSTM models with attention further improved recall of subtle subjective cues, but these methods required large amounts of annotated data in each target language.

Transformer-based pretrained models revolutionized cross-lingual transfer. Devlin et al. (2019) introduced BERT, and its multilingual variant (`mBERT`) showed strong zero-shot performance across over 100 languages. Subsequent work by Barnes et al. (2020) specifically fine-tuned mBERT for cross-lingual subjectivity detection, demonstrating consistent gains over BiLSTM baselines [9].

Despite these advances, subjectivity classification in low-resource and unseen languages remains challenging. Our work builds on the strengths of mBERT's cross-lingual embeddings and introduces a lightweight post-processing module for English to refine low-confidence predictions using interpretable lexical features.

3. Dataset

All experiments in this study were conducted using the annotated dataset provided by the *CheckThat! 2025 Lab: Task 1 – Subjectivity Detection* [10, 11, 12, 13]. The dataset is multilingual and spans five primary training languages—English, Italian, German, Bulgarian, and Arabic—with additional zero-shot evaluation on five other languages, namely Ukrainian, Romanian, Greek, Polish, and a multilingual mix. Each training language includes data splits for training, development, and development-test, while the evaluation phase was performed on an unseen test set. The dataset consisted of sentences extracted from news articles in multiple languages which were annotated as either subjective (SUBJ) or objective (OBJ) as per the guidelines developed by Antici et al. [14] that defines a sentence to be subjective if its content is based on or influenced by personal feelings, tastes, or opinions. Otherwise, the sentence is objective. Table 1 presents the dataset statistics for each of the five primary languages. In general, the dataset exhibits class imbalance where number of objective sentences is more than that of subjective ones across majority of the languages.

Table 1
Distribution of objective (OBJ) and subjective (SUBJ) sentences across train, dev, and dev-test splits for each of the five primary languages.

Language	Train		Dev		Dev-Test	
	OBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ
English	532	298	222	240	362	122
Italian	1231	382	490	177	377	136
German	492	308	317	174	226	111
Bulgarian	406	323	175	139	143	107
Arabic	1391	1055	266	201	425	323

In addition to Table 1, the English test set contained 300 sentences, the Italian set had 299, the German test set included 347 sentences, and Arabic had the largest test set with 1036 senetnces. Notably, no official test data was released for Bulgarian.

Beyond the core training languages, the test set also included data for evaluating zero-shot performance. Specifically, the Ukrainian test set contained 297 sentences, the Romanian set had 206, the

Greek test set included 284, and Polish comprised 351 test instances. Furthermore, a multilingual test set with 1982 sentences drawn from a diverse mix of languages was provided to assess cross-lingual generalization capabilities.

4. Methodology

This section describes the methodologies we employed for subjectivity classification. Given an input sentence T, our goal was to determine whether T expressed a subjective viewpoint (SUBJ) or an objective statement (OBJ).

4.1. Text Preprocessing

All raw sentences were first cleaned and normalized. We removed escape characters such as '\n' (newline) and '\t' (tab), and converted any Unicode characters to their ASCII equivalents. After cleaning, each sentence was tokenized into a sequence of word tokens $[k_1, k_2, \ldots, k_n]$ using the NLTK tokenizer.

4.2. Textual Analysis and Lexical Feature Extraction

For the English monolingual subtask, we defined two interpretable, lexicon-based features to refine low-confidence BERT predictions.

4.2.1. SentiWordNet Subjective Score

Inspired by Esuli and Sebastiani, we computed a subjective score for each sentence based on SentiWordNet synsets. Each token was POS-tagged and mapped to a WordNet tag; for each unique lemma, we averaged the sum of positive and negative scores across its synsets. The sentence-level score was obtained by summing these averages:

$$\text{Score}_{\text{SWN}}(T) \;=\; \sum_{\ell \in \mathcal{L}(T)} \frac{1}{|\mathcal{S}(\ell)|} \sum_{s \in \mathcal{S}(\ell)} \big(\text{pos\_score}(s) + \text{neg\_score}(s)\big),$$

where $\mathcal{L}(T)$ is the set of lemmatized tokens in T and $\mathcal{S}(\ell)$ are the SentiWordNet synsets for lemma ℓ. Sentences with $\text{Score}_{\text{SWN}} \geq 1.615$ were considered subjective.

4.2.2. Combined Lexical-Clues and Opinion-Lexicon Count

We counted occurrences of lexical clues that often reflect subjectivity, such as "perhaps", "I believe", "wonderful", "wow", "my pleasure", "sorry", and others. Some of these phrases were adapted from the human-phrase list proposed in [15], which showed that such expressions are more commonly used by humans than AI which can be linked with personal opinions and subjective content. Since our task focused on identifying subjective language, these clues were suitable indicators.

In addition, we included extra phrases that matched our task requirements — specifically hedge words (e.g., "maybe", "likely", "suppose") and sarcastic or emotional expressions (e.g., "yeah right", "totally", "as if"), as these also signal subjectivity. All the lexical clues that we considered are provided in Appendix A. To further strengthen this feature, we also counted sentiment-bearing words from the Hu and Liu Opinion Lexicon. Let Φ be the set of human-phrases and Ω the set of opinion-lexicon words. For each sentence T, we computed

$$\text{Count}(T) \;=\; \sum_{\phi \in \Phi} \text{freq}(\phi, T) \;+\; \sum_{\omega \in \Omega} \text{freq}(\omega, T).$$

Sentences with $\text{Count}(T) \geq 4$ were labeled as subjective.

4.2.3. Threshold Selection Methodology

To choose the best threshold for each feature, we experimented with several values of threshold by performing binary classification with each value on the English training dataset and checked which one gave the highest classification accuracy. For the SentiWordNet Subjective Score, we tried thresholds such as 1.5, 1.6, 1.615, 1.625, 1.65, 1.7 etc. Among these, 1.615 gave the best accuracy in separating subjective and objective sentences, so we used it as the final threshold.

For the Combined Lexical-Clues and Opinion-Lexicon Count, we experimented with thresholds like 2, 3, 4, 5 etc. A threshold of 4 gave the best balance — it correctly identified most of the subjective sentences without misclassifying too many objective ones.

5. Framework Development

The overall framework is shown in Figure 1 where SentiW denotes the SentiWordNet Subjective Score feature and the Feat denotes Combined Lexical-Clues and Opinion-Lexicon Count feature. We built our classifier on top of the transformer-based `bert-base-multilingual-cased` (mBERT) model [3]. We selected mBERT because it was pretrained on the Wikipedia dumps of 104 languages, yielding strong contextual representations across diverse linguistic families, and because it has demonstrated robust zero-shot transfer capabilities without target-language fine-tuning [16]. This multilingual pretraining allows mBERT to leverage shared subword vocabularies and syntactic patterns, making it particularly suitable for our nine-subtask setting that spans both high- and low-resource languages. Each input sentence was tokenized using the mBERT tokenizer to produce `input_ids` and `attention_mask` tensors. These tensors were passed through the mBERT encoder, yielding a sequence of hidden states

$$\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}.$$

We then selected the hidden state corresponding to the special `[CLS]` token, denoted $\mathbf{h}_{[\mathrm{CLS}]}$, as the aggregate representation of the entire sentence.

Classification: The vector $\mathbf{h}_{[\mathrm{CLS}]}$ was fed into the built-in classification head of mBERT, which applies dropout (with the default probability of 0.1) and a linear layer to produce two logits (z_0, z_1). These logits were converted into class probabilities via the softmax function:

$$P_i = \frac{\exp(z_i)}{\exp(z_0) + \exp(z_1)}, \quad i \in \{0, 1\},$$

and the final label was chosen as

$$\hat{y} = \arg\max_{i \in \{0,1\}} P_i,$$

where $\hat{y} = 1$ indicates a subjective sentence and $\hat{y} = 0$ indicates an objective sentence.

6. Training

We used the training and development splits provided by the CheckThat! 2025 Task 1 organizers, further stratifying each split into 90% for training and 10% for validation. We fine-tuned the model for two epochs with a per-device batch size of eight, using standard cross-entropy loss function and AdamW optimizer [17] with a learning rate of 2×10^{-5}. We configured the Trainer to evaluate at the end of each epoch, log training progress every 100 steps, and save model checkpoints after each epoch. All experiments used a fixed random seed of 42 for reproducibility.

7. Post-processing

After training, we applied our feature-based post-processing only to the English monolingual test predictions with confidence less than 0.7 to recompute its label by evaluating the two lexical features

described in Section 4.2. Specifically, each sentence with maximum predicted probability $\max_i P_i < 0.7$ was assigned SUBJ only if both its SentiWordNet subjective score exceeded 1.615 and its combined lexical-clues plus opinion-lexicon count exceeded 4; otherwise, it was labeled OBJ. This strategy leveraged interpretable linguistic cues to correct uncertain model outputs.

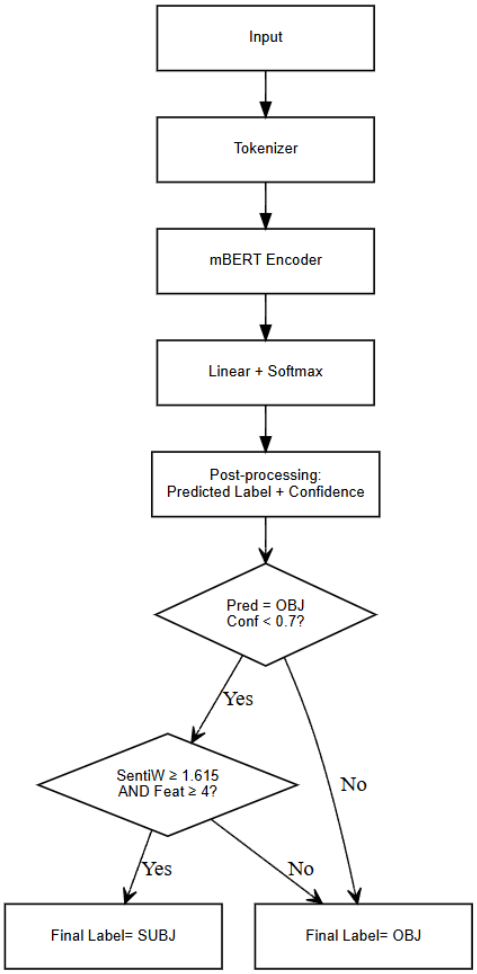

Figure 1: Framework diagram of mBERT model with post-processing

8. Experimental Setup

All the experiments were accomplished using Python libraries such as `pandas`, `nltk`, `datasets` and `scikit-learn`. The transformer-based mBERT model was trained and evaluated using HuggingFace's `transformers` library with support from `PyTorch`. The development and execution of the model pipeline, including training and inference, were carried out entirely in the Kaggle environment equipped with NVIDIA Tesla T4 dual GPUs.

For English-specific post-processing, we used additional Python modules such as `nltk`'s `sentiwordnet`, `wordnet`, `opinion_lexicon`, and `scipy`'s `softmax` function to calculate lexical features. To evaluate the performance of the proposed subjectivity classification framework, macro-F1 scores were calculated for the exact test data provided by the organizers of the CheckThat! 2025 Task-1.

9. Results

We evaluated the results for the nine subtasks under the CheckThat! 2025 Task 1 using the fine-tuned mBERT model. The task comprises three evaluation settings: monolingual, multilingual, and zero-shot. The monolingual subtasks include Arabic, English, German, Italian, and Bulgarian; the multilingual setting involves joint training and testing on all five training languages; and the zero-shot setting tests on unseen languages—Greek, Polish, Ukrainian, and Romanian—after training on the multilingual corpus.

The macro-averaged F1 scores and the team rank of the mBERT classifier for each subtask are provided in Table 2.

Table 2
Macro-F1 results of fine-tuned mBERT and team ranks across nine subtasks (without post-processing).

Language	Setting	Macro-F1	Data Used	Rank
Arabic	Monolingual	0.545	Test	14
English	Monolingual	0.693	Test	-
English	Monolingual	0.730	Dev-Test	-
Italian	Monolingual	0.699	Test	10
German	Monolingual	0.736	Test	9
Multilingual	Multilingual	0.654	Test	10
Polish	Zero-shot	0.560	Test	11
Ukrainian	Zero-shot	0.580	Test	11
Romanian	Zero-shot	0.744	Test	8
Greek	Zero-shot	0.435	Test	8

Among all subtasks, the best performance was achieved on Romanian in the zero-shot setting with a macro-F1 of 0.744. In contrast, the lowest result was obtained for the Greek zero-shot setting (0.435), indicating a possible gap in representation coverage for low-resource languages with limited overlap in the multilingual vocabulary. Our team achieved a strong rank of 8 in this setting in Romanian, while we matched the same rank in Greek despite its low score. Romanian's strong result may be attributed to its linguistic similarity with Italian and the presence of shared subword units in mBERT's vocabulary, allowing better transfer from the multilingual training set. On the other hand, Greek, being morphologically rich and having limited lexical overlap with the training languages, likely suffered from poor representation alignment in the mBERT embedding space.

In case of German and Italian languages, the model performed robustly, with macro-F1 scores of 0.736 and 0.699 respectively therefore achieving a team rank of 9 and 10 respectively, likely due to sufficient training data and the presence of high-resource status in the original pretraining corpus of mBERT.

English, despite being one of the primary pretraining languages for mBERT, showed slightly lower performance (0.693) on test dataset whereas it performed slightly better on development-test dataset (0.730). The reason for low performance on English can be possibly due to the inherent difficulty of distinguishing subjective expressions from objective ones within its dataset. The rank achieved on the test set after post-processing is discussed later in Table 4.

Arabic, with a macro-F1 of 0.545, underperformed relative to other monolingual settings, which may stem from its complex morphology, right-to-left structure, and limited overlap with other Latin-based training languages used in the multilingual setup. The model ranked 14th in this subtask.

The multilingual setting, which involved joint training and testing on all five training languages, resulted in a macro-F1 score of 0.654 with a team rank of 10. While this value was slightly lower than some monolingual track performances, it confirmed our fine-tuned mBERT's ability to generalize reasonably well when trained on mixed-language data, balancing performance across a broader linguistic spectrum.

Result after post-processing: To further enhance classification in the English monolingual setting, we applied a post-processing step on top of the mBERT classifier. The post-processing was applied

selectively to predictions with confidence lower than a threshold value θ, using lexical features described earlier in Section 4.5.

Table 3 presents the macro-F1 scores on the English development-test split across multiple confidence scores such as 0.99,0.95,0.90,0.85,0.75,0.70 and 0.65.

Table 3
Performance of mBERT model with post-processing on English subtask (dev-test).

Confidence(conf)	Macro-F1	Data Used
conf < 0.99	0.6581	Dev–Test
conf < 0.95	0.6749	Dev–Test
conf < 0.90	0.7054	Dev–Test
conf < 0.85	0.7283	Dev–Test
conf < 0.80	0.7478	Dev–Test
conf < 0.75	0.7562	Dev–Test
conf < 0.70	0.7575	Dev–Test
conf < 0.65	0.7492	Dev–Test

From Table 3, it is evident that the post-processing technique consistently improved the macro-F1 scores compared to the baseline mBERT model, which scored a macro-F1 of 0.693 without any post-hoc adjustments.

Thresholds below `conf` < 0.65 or above `conf` < 0.75 resulted in diminishing macro-F1 gains, indicating that the post-processing mechanism is only effective within this intermediate confidence range. Based on the analysis of Table 3, the best performance was achieved at a confidence threshold of `conf` < 0.70, where the macro-F1 peaked at 0.7575 due to which the value `conf` < 0.70 was selected as the optimal operating point. This threshold offered a balanced trade-off between overall performance and robustness against uncertain predictions.

Using this chosen threshold, we applied the post-processing pipeline on the English subtask test set to assess its effectiveness beyond the development data. The results before and after applying the post-processing technique along with the team rank in this subtask are summarized in Table 4.

Table 4
Effect of post-processing on English test set and team rank in English subtrack.

Language	Setting	**Before**	**After**	Data Used	Rank
English	Monolingual	0.693	0.733	Test	8

As shown in Table 4, the macro-F1 score improved from 0.693 to 0.733 on the English test data after post-processing achieving a remarkable team rank of 8 among all other teams. This result reinforces the benefit of applying interpretable, lexical-based heuristics in combination with transformer predictions, particularly when dealing with borderline or ambiguous classifications.

Overall, the proposed mBERT-based subjectivity classification framework demonstrated promising performance across diverse linguistic settings and was further strengthened by lightweight post-processing enhancements that addressed model uncertainty in a principled and explainable manner.

10. Conclusion

We presented a compact mBERT-based framework for multilingual subjectivity classification and evaluated it across monolingual, multilingual, and zero-shot settings in CheckThat! 2025 Task 1. The approach delivered strong performance in high-resource languages—German (0.736 macro-F1 score), Italian (0.699 macro-F1 score) and English (0.693 macro-F1 score)—and achieved notable zero-shot transfer for Romanian (0.744 macro-F1 score), while revealing challenges for morphologically rich languages such as Greek and Arabic. The multilingual model attained a macro-F1 score of 0.654,

demonstrating reasonable cross-language generalization when trained on mixed-language data. A lightweight post-processing step further boosted English performance, improving macro-F1 score from 0.693 to 0.733 on the test set. These results demonstrated the value of combining transformer representations with interpretable lexical heuristics for robust, explainable subjectivity detection across diverse languages.

11. Limitations

Despite encouraging results, our work presents several limitations. First, performance varied notably across languages, especially in zero-shot settings. The poor result on Greek suggests that mBERT's pretraining does not equally benefit all languages, particularly those with distinct orthographies or limited token overlap with the training set. This calls for future research into adaptive multilingual pretraining or low-resource augmentation strategies to close the performance gap.

Second, the post-processing module was designed specifically for English, relying on lexical cues and syntactic constructs that may not directly transfer to other languages. While the module improved both overall and subjective-class accuracy, extending it to non-English settings would require developing language-specific rules that respect local morphological and syntactic properties.

Lastly, the classification was conducted at the sentence level without considering discourse or contextual information. Integrating document-level context or inter-sentence dependencies could help in resolving subtler subjectivity distinctions that span across sentences. Future work could also explore hybrid architectures that blend LLMs with graph-based discourse models for deeper semantic understanding.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* (OpenAI) for the sole purpose of **Paraphrase and reword**. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.
Tools and services: ChatGPT (OpenAI)
Tools' contributions (GenAI Usage Taxonomy): Paraphrase and reword.
No generative AI system is listed as an author, and all core scientific contributions (problem formulation, experiments, analyses, and conclusions) were performed solely by the human authors.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019) 4171–4186.

[2] A. Conneau, G. Lample, Unsupervised cross-lingual representation learning at scale, in: ACL, 2019, pp. 8440–8451.

[3] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[4] A. Esuli, F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in: LREC, 2006, pp. 417–422.

[5] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 168–177. URL: https://doi.org/10.1145/1014052.1014073. doi:10.1145/1014052.1014073.

[6] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of ACL, 2004, pp. 271–278.

[7] J. M. Wiebe, R. Bruce, T. O'Hara, Recognizing subjective sentences: A subjectivity lexicon approach, in: Proceedings of the ACL Workshop on Text Classification, 2002, pp. 45–52.
[8] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of EMNLP, 2014, pp. 1746–1751.
[9] J. Barnes, Y. Zhao, Cross-lingual subjectivity detection with multilingual bert, in: Proceedings of EMNLP, 2020, pp. 1234–1245.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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/.
[15] U. Jawaid, R. Roy, P. Pal, S. Debnath, D. Das, S. Bandyopadhyay, Human vs machine: An automated machine-generated text detection approach, in: S. Lalitha Devi, K. Arora (Eds.), Proceedings of the 21st International Conference on Natural Language Processing (ICON), NLP Association of India (NLPAI), AU-KBC Research Centre, Chennai, India, 2024, pp. 215–223. URL: https://aclanthology.org/2024.icon-1.24/.
[16] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, Proceedings of NAACL-HLT (2019) 4996–5001.
[17] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.

A Appendix: Lexical-Clues

The following list of Lexical Clues was used in our post-processing heuristic to help distinguish subjective text. These phrases include sarcastic words, hedging expressions, personal references, and emotional cues commonly found in subjective content.

1. "therefore"
2. "however"
3. "etc"
4. "whatever"
5. "?"
6. "maybe"

7. “perhaps”
8. “seems”
9. “appears”
10. “likely”
11. “unlikely”
12. “could”
13. “might”
14. “would”
15. “suggest”
16. “suppose”
17. “probably”
18. “supposedly”
19. “preferably”
20. “but”
21. “yeah right”
22. “totally”
23. “as if”
24. “great”
25. “wonderful”
26. “oh sure”
27. “wow”
28. “just”
29. “awesome”
30. “!”
31. “...”
32. “my god”
33. “shit”
34. “what”
35. “damn”
36. “my pleasure”
37. “sorry”
38. “sorry to say”
39. “thank you”
40. “consequently”
41. “meanwhile”
42. “in addition”
43. “furthermore”
44. “moreover”
45. “nevertheless”
46. “nonetheless”
47. “in the same way”
48. “let”
49. “suppose”
50. “imagine”
51. “think about it”
52. “think about that”
53. “honestly”
54. “precisely”
55. “your”
56. “you”
57. “mine”
58. “my”
59. “i”
60. “yourself”
61. “i think”
62. “just imagine”
63. “i believe”
64. “what do you”
65. “when do you”
66. “have you”