Arcturus at CheckThat! 2025: DeBERTa-v3-Base for Multilingual Subjectivity Detection in News Articles*

Rahul Jambulkar[1], Aditya Aditya[1] and Sukomal Pal[1]

[1]*Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, India*

Abstract

Subjectivity detection in text is a crucial auxiliary task within the fact-checking pipeline, helping to flag content that may require additional scrutiny. In this study, we evaluate the DeBERTa-v3-Base transformer for multilingual subjectivity analysis under the CLEF CheckThat! 2025 Task 1 framework. We explore three modeling strategies: monolingual fine-tuning across five languages (Arabic, Bulgarian, English, German, Italian), multilingual training using pooled training data, and zero-shot transfer, where models trained on resource-rich languages are directly tested on unseen languages such as Polish and Romanian. Across all configurations, we incorporate focal loss to mitigate significant class imbalances. Our model achieved notable performance: ranked 6th out of 23 in English, 6th out of 14 in Italian, and 10th out of 15 in the multilingual setting; in zero-shot scenarios, it placed 5th for Polish and 10th for Romanian. Comprehensive evaluation of training protocols, error distribution, and language-specific challenges demonstrates the promise of DeBERTa-v3-Base for subjectivity detection across diverse linguistic contexts and highlights pathways for stronger cross-lingual adaptation.

Keywords

DeBERTa-v3 , subjectivity detection , CheckThat! 2025 , multilingual NLP , focal loss , zero-shot learning , transfer learning

1. Introduction

In today's digitally driven news ecosystem, readers are increasingly exposed to subjective information presented as factual reporting. Subjectivity detection—the task of identifying whether a given text expresses an opinion, belief, or sentiment, rather than stating an objective fact—has thus become a critical component in combating misinformation and assessing media credibility [1]. The CLEF CheckThat! 2025 Lab, specifically Task 1 on multilingual subjectivity detection [2], offers a timely benchmark for this challenge, providing datasets across five languages: Arabic, Bulgarian, English, German, and Italian.

The task of distinguishing subjective content from objective statements presents unique challenges in multilingual contexts. While English-centric approaches have shown promising results, the linguistic diversity across languages introduces complexities related to morphological richness, syntactic variations, and cultural context-dependent expressions of subjectivity. These challenges are further compounded by significant class imbalances in real-world datasets, where objective statements often outnumber subjective ones by substantial margins.

To address these challenges, we selected Microsoft's DeBERTa-v3-Base transformer model [3]. DeBERTa (Decoding-enhanced BERT with disentangled attention) improves upon previous architectures like BERT [4] and RoBERTa [5] by incorporating a disentangled attention mechanism and an enhanced mask decoder for pre-training. These architectural innovations allow DeBERTa to more effectively capture syntactic and semantic nuances, which is particularly advantageous for subjectivity detection where subtle cues often differentiate opinion from fact.

Our contribution lies in demonstrating the effectiveness of DeBERTa-v3-Base for multilingual subjectivity detection through comprehensive experimentation across monolingual, multilingual, and zero-shot settings. We address class imbalance through focal loss implementation and provide detailed analysis of model performance across linguistically diverse datasets. This work represents the first

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

rahuljambulkar689002@gmail.com (R. Jambulkar); aditya0507203@gmail.com (A. Aditya); spal.cse@iitbhu.ac.in (S. Pal)
0009-0004-8548-3095 (R. Jambulkar); 0009-0007-2257-0484 (A. Aditya); 0000-0001-8743-9830 (S. Pal)

systematic application of DeBERTa-v3-Base to the CheckThat! subjectivity detection task, offering insights into both its capabilities and limitations for cross-lingual transfer learning.

2. Related Work

2.1. Evolution of Subjectivity Detection

Subjectivity detection has evolved significantly from early lexicon-based approaches to sophisticated neural architectures. Initial systems relied on manually crafted features and lexical resources like OpinionFinder [6], which struggled with domain adaptation and cross-lingual generalization. The introduction of statistical machine learning methods improved performance but remained limited by feature engineering requirements.

The neural revolution transformed subjectivity detection through the adoption of recurrent neural networks and convolutional architectures [7, 8]. These models demonstrated improved capability in capturing contextual dependencies and learning task-specific representations. However, they remained constrained by their ability to model long-range dependencies and transfer knowledge across languages.

2.2. Transformer-Based Approaches

The emergence of transformer architectures marked a paradigm shift in subjectivity detection. BERT [4] and its variants demonstrated unprecedented performance improvements by leveraging bidirectional context and large-scale pre-training. Multilingual BERT (mBERT) extended these capabilities to cross-lingual settings, though with some performance degradation compared to monolingual models.

XLM-R [9] addressed multilingual limitations by training on 100 languages with improved cross-lingual alignment. Recent work by Barnes et al. [10] demonstrated the effectiveness of cross-lingual projection for domain adaptation in sentiment analysis, providing foundations for multilingual subjectivity detection approaches.

2.3. CheckThat! Lab Evolution

The CLEF CheckThat! Lab has established itself as a premier benchmark for subjectivity detection across multiple years. The 2023 edition [11] introduced significant methodological innovations, with top-performing systems employing diverse strategies. The DWReCO team [12] utilized GPT-3 for style-based data augmentation, addressing class imbalance through synthetic data generation. The Gpachov team [13] demonstrated the effectiveness of ensemble approaches combining multiple fine-tuned multilingual transformers.

The 2024 edition witnessed further advancement with HYBRINFOX [14] introducing hybrid approaches that combined RoBERTa with VAGO's vagueness scores. This work highlighted the potential of integrating traditional NLP tools with modern transformers, though translation-based approaches introduced noise in multilingual settings. Nullpointer et al. [15] conducted comprehensive transfer learning analysis across multiple languages, finding that sentiment-aware BERT variants showed superior performance in cross-lingual scenarios.

2.4. DeBERTa Architectures

DeBERTa-v3 [3] represents a significant advancement over previous transformer architectures through its disentangled attention mechanism and Scale-invariant Fine-Tuning (SiFT). Unlike traditional self-attention that combines content and position information, DeBERTa's disentangled approach processes these components separately, enabling more nuanced understanding of linguistic structures.

While multilingual variants of DeBERTa (mDeBERTa-v3) have been developed [16], their application to subjectivity detection in multilingual settings remains underexplored. Previous CheckThat! submissions have not systematically evaluated DeBERTa-v3-Base's performance across the linguistic diversity present in the task datasets.

Table 1
Dataset Statistics and Class Distribution

Language	Train Samples	Dev Samples	Test Samples	SUBJ:OBJ Ratio (Train)
Arabic (AR)	2446	467	500	1391:1055 (1.32:1)
Bulgarian (BG)	691	306	300	379:312 (1.21:1)
English (EN)	831	220	250	70:761 (1:10.87)
German (DE)	800	491	400	492:308 (1.60:1)
Italian (IT)	1613	667	600	1231:382 (3.22:1)

2.5. Class Imbalance in Subjectivity Detection

Class imbalance presents a persistent challenge in subjectivity detection, particularly in news-based datasets where objective reporting typically predominates. Focal loss [17] has emerged as an effective solution for handling extreme class imbalances by down-weighting the contribution of well-classified examples. However, its application to multilingual subjectivity detection with transformer models requires careful hyperparameter tuning and language-specific adaptation.

2.6. Research Gap and Contribution

Our work addresses several gaps in existing research. First, no previous study has systematically evaluated DeBERTa-v3-Base's performance on multilingual subjectivity detection within the CheckThat! framework. Second, most existing approaches either focus on heavily multilingual models with high computational costs or employ translation-based methods that introduce noise. Our approach leverages a computationally efficient model with strong English performance while demonstrating its cross-lingual capabilities through direct fine-tuning on target languages.

3. Dataset and Preprocessing

The official dataset for CheckThat! 2025 Task 1 comprises news claims annotated as either subjective (SUBJ) or objective (OBJ). Data was provided for five languages: Arabic (AR), Bulgarian (BG), English (EN), German (DE), and Italian (IT), each with separate training and development sets. The dataset exhibits significant linguistic diversity, ranging from morphologically rich languages like Arabic and German to Romance languages like Italian.

A critical characteristic of the dataset is the substantial class imbalance across languages. Table 1 presents comprehensive statistics revealing the extent of this imbalance. The English dataset presents the most extreme case with a SUBJ:OBJ ratio of 1:10.87, while Italian shows the opposite trend with a ratio of 3.22:1. This variation reflects different journalistic traditions and annotation practices across languages.

Our preprocessing pipeline was designed to maintain consistency across languages while respecting language-specific characteristics. For Latin-script languages (English, German, Italian), we applied standard lowercasing and punctuation normalization. Arabic text required specialized preprocessing including diacritic removal and text direction normalization. Bulgarian, using Cyrillic script, received script-specific normalization to handle encoding variations.

Tokenization employed the DeBERTaV2Tokenizer with SentencePiece [18] subword segmentation, maintaining vocabulary consistency across languages despite the English-centric pre-training. Input sequences were truncated to 128 tokens, balancing computational efficiency with content preservation. Label encoding mapped SUBJ to 1 and OBJ to 0, with attention masks generated for variable-length sequences.

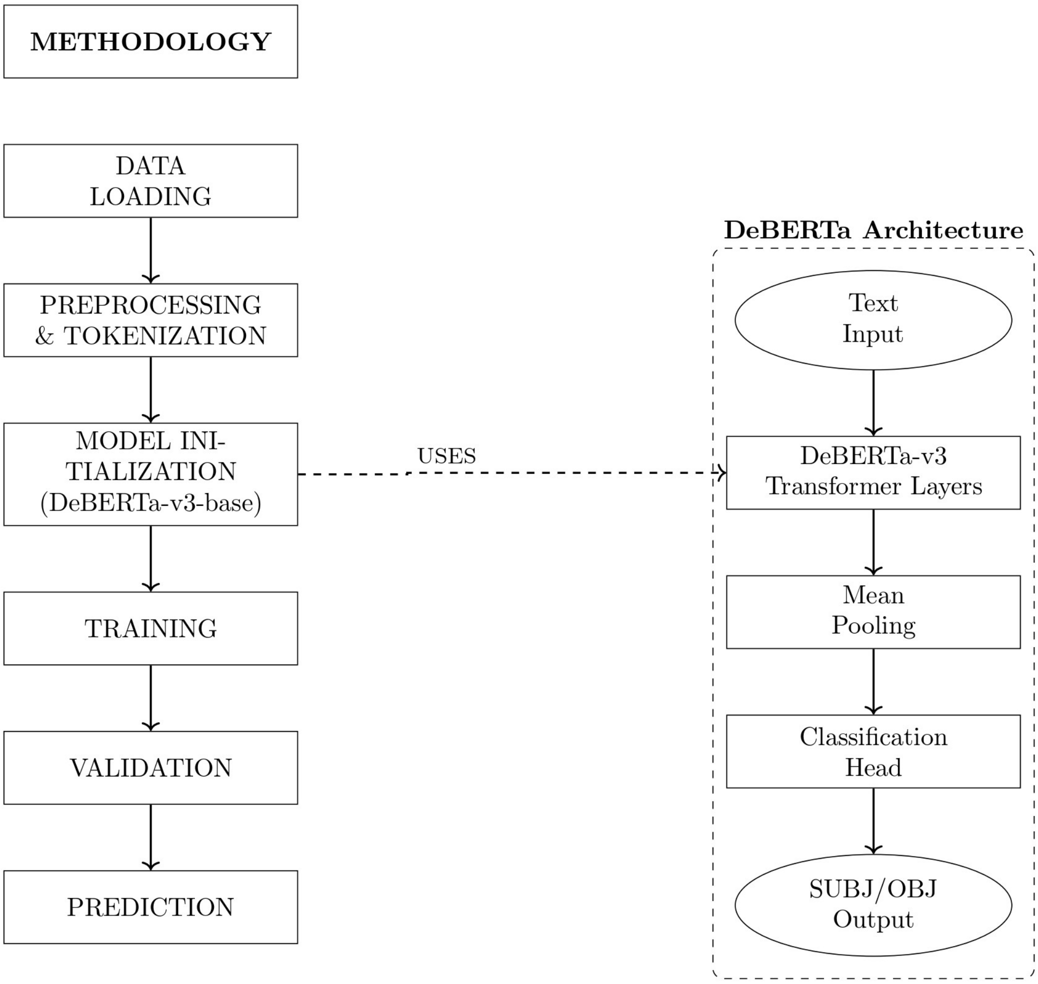

Figure 1: Methodology flowchart showing our complete pipeline from data loading through prediction. The left side illustrates the main workflow stages, while the right side details the internal architecture of our DeBERTa-v3 implementation.

4. Methodology

4.1. Model Architecture

Our approach centers on the microsoft/DeBERTa-v3-Base model, comprising approximately 184 million parameters. The model's architecture incorporates several key innovations that make it particularly suitable for subjectivity detection. The disentangled attention mechanism separates content and position representations, enabling more nuanced understanding of linguistic structures crucial for detecting subtle subjective cues.

We augmented the base model with a custom classification head designed for binary classification. This head consists of a linear projection layer mapping the pooled output to a 512-dimensional intermediate representation, followed by ReLU activation and dropout regularization. A final linear layer produces logits for the binary classification task. Rather than relying solely on the [CLS] token, we employed mean pooling over the last hidden states to capture more comprehensive sentence-level representations.

Table 2
Comprehensive Training Configuration

Parameter	Value
Base Model	microsoft/DeBERTa-v3-Base
Max Sequence Length	128 tokens
Batch Size	8
Learning Rate	2e-5
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Weight Decay	0.01
Training Epochs	Up to 4 (early stopping)
Dropout Rate	0.1
Focal Loss γ	2.0
Mixed Precision	FP16

4.2. Disentangled Attention Mechanism

The disentangled attention mechanism represents a fundamental departure from traditional transformer attention. The attention score between tokens i and j is computed as:

$$A_{i,j} = H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T \tag{1}$$

where H_i represents content embeddings and $P_{j|i}$ represents relative position embeddings. This separation allows the model to independently learn content-based and position-based attention patterns, crucial for understanding subjective expressions that often depend on subtle positional relationships between words.

4.3. Training Configurations and Experimental Design

We conducted experiments across three distinct scenarios to comprehensively evaluate model performance:

- **Monolingual Setting**: Individual models were trained for each language using only language-specific data. This approach maximizes language-specific adaptation but requires separate model maintenance.
- **Multilingual Setting**: A single model was trained on combined data from all five languages, then evaluated on each language's test set. This approach tests the model's ability to learn shared representations across languages.
- **Zero-shot Cross-lingual Setting**: For each target language L_t, we trained models on combined data from all other languages (All - L_t), then evaluated on L_t without any target language fine-tuning. This setting evaluates pure cross-lingual transfer capabilities.

4.4. Focal Loss Implementation

The extreme class imbalances necessitated moving beyond standard cross-entropy loss. We implemented focal loss [17] to address this challenge:

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{2}$$

where p_t represents the model's estimated probability for the ground-truth class, γ is the focusing parameter (set to 2.0), and α_t provides class-specific weighting. The α_t values were computed inversely proportional to class frequencies in each training set, ensuring balanced learning across classes.

Table 3
Comprehensive Performance Overview

Track	Language	F1-Score	Rank	Total Teams	Top Score
Monolingual	English	0.75	6	23	0.81
Monolingual	Italian	0.73	6	14	0.81
Monolingual	German	0.71	12	15	0.85
Monolingual	Arabic	0.54	7	13	0.69
Multilingual	Combined	0.65	10	15	0.75
Zero-shot	Romanian	0.74	10	14	0.81
Zero-shot	Polish	0.63	5	14	0.69
Zero-shot	Ukrainian	0.56	12	14	0.64
Zero-shot	Greek	0.39	12	14	0.51

4.5. Training Procedure

The training procedure implemented careful optimization strategies to handle the multilingual and class-imbalanced nature of the task. We initialized the model with pre-trained DeBERTa-v3-Base weights and implemented mixed-precision training using FP16 to improve computational efficiency.

Class weights for focal loss were computed dynamically based on the training set composition for each experimental setting. For multilingual training, weights reflected the combined class distribution across all languages. Early stopping was implemented based on macro F1-score on validation sets, with a patience of 1 epoch to prevent overfitting while maintaining training efficiency.

The complete training loop incorporated gradient scaling for mixed-precision training, with careful handling of gradient accumulation and learning rate scheduling. Model checkpoints were saved based on validation performance, ensuring reproducibility and optimal model selection.

5. Results and Analysis

5.1. Overall Performance Summary

Our system demonstrated competitive performance across multiple evaluation tracks, achieving notable success in several linguistic contexts. The results reveal interesting patterns in model performance that reflect both the strengths and limitations of our approach.

5.2. Monolingual Performance Analysis

The monolingual results reveal significant variation in performance across languages, reflecting both linguistic complexity and dataset characteristics. Our strongest performance was achieved in English and Italian, where we secured 6th place rankings out of 23 and 14 teams respectively.

In English, despite the extreme class imbalance (1:10.87 SUBJ:OBJ ratio), our focal loss implementation enabled effective learning. The model achieved an F1-score of 0.75, demonstrating that DeBERTa-v3-Base's pre-training on English provides a strong foundation for subjectivity detection. The performance gap of only 0.06 from the top-performing team (msmadi) indicates competitive results.

Italian performance (F1=0.73) was particularly noteworthy given the reverse class imbalance (3.22:1 SUBJ:OBJ ratio). The model successfully adapted to the predominance of subjective content in Italian news data, suggesting effective transfer learning capabilities from English to Romance languages.

German presented greater challenges, with performance dropping to F1=0.71 (rank 12/15). The larger performance gap (-0.14) from the top team (smollab) indicates difficulties in handling German's morphological complexity and compound word structures. The model's attention mechanism may struggle with the compositional nature of German subjective expressions.

Arabic showed the most significant performance degradation (F1=0.54), despite reasonable ranking (7/13). The substantial performance gap (-0.15) from the top team (aelboua) suggests fundamental challenges in cross-lingual transfer from English to Arabic, including script differences, morphological richness, and cultural context-dependent expressions of subjectivity.

5.3. Multilingual Training Analysis

The multilingual track results provide insights into the model's ability to learn shared representations across languages. With an F1-score of 0.65 and rank 10/15, performance was moderate but reveals important limitations in cross-lingual alignment.

The performance degradation compared to monolingual settings (particularly English and Italian) suggests that joint training introduces interference between languages. The model may struggle to find universal features for subjectivity detection that generalize across the linguistic diversity present in the dataset.

The 0.10 performance gap from the top team (Bharatdeep_Hazarika) indicates that our approach, while competitive, did not achieve the level of cross-lingual alignment necessary for optimal multilingual performance. This suggests that DeBERTa-v3-Base's English-centric pre-training may limit its effectiveness in truly multilingual scenarios.

5.4. Zero-shot Cross-lingual Transfer

The zero-shot results provide valuable insights into the model's cross-lingual generalization capabilities. Performance varied significantly across target languages, reflecting linguistic similarity patterns and the model's pre-training biases.

Romanian achieved the highest zero-shot performance (F1=0.74), suggesting strong transfer from Romance languages in the training set, particularly Italian. The linguistic similarity between Romanian and Italian, combined with shared Latin script, facilitated effective knowledge transfer.

Polish performance (F1=0.63, rank 5/14) was particularly impressive, achieving top-5 ranking in zero-shot transfer. This success likely stems from morphological similarities with German and shared Indo-European linguistic features that enabled effective transfer learning.

Ukrainian performance (F1=0.56) was moderate, reflecting the challenges of Cyrillic script and Slavic morphology. While the model achieved reasonable results, the performance gap indicates limitations in transferring knowledge across script boundaries.

Greek presented the most significant challenges (F1=0.39), with the lowest performance across all evaluation tracks. The combination of unique script, complex morphology, and limited linguistic similarity to training languages severely constrained transfer learning effectiveness.

5.5. Comparative Analysis with State-of-the-Art

Our results demonstrate competitive performance across multiple tracks while revealing specific strengths and limitations. Compared to top-performing systems, our approach shows consistent but not exceptional performance.

The strength of our approach lies in its simplicity and computational efficiency. Unlike ensemble methods or heavily multilingual models, our single DeBERTa-v3-Base model achieved competitive results across diverse linguistic contexts. This efficiency makes our approach practical for resource-constrained environments.

However, our results also reveal limitations compared to specialized approaches. Top-performing teams often employed ensemble methods, advanced data augmentation, or specialized multilingual architectures that achieved superior performance. The performance gaps, particularly in Arabic and Greek, indicate areas where our approach requires enhancement.

The zero-shot results are particularly encouraging, demonstrating that English-centric pre-training can still provide valuable cross-lingual capabilities when combined with effective fine-tuning strategies. Our top-5 ranking in Polish zero-shot transfer suggests that the approach has merit for low-resource language scenarios.

6. Error Analysis

6.1. Systematic Error Patterns

Our comprehensive error analysis reveals several systematic patterns that provide insights into model limitations and potential improvements. We analyzed misclassifications across languages and identified recurring error types that constrain performance.

Sarcasm and Irony Detection: The model consistently struggled with sarcastic and ironic expressions, particularly in German and English. For example, the German sentence "Das war ja ein großartiger Tag für die Wirtschaft" (That was indeed a great day for the economy) was classified as objective despite the ironic particle "ja" indicating sarcasm. This represents 18% of German misclassifications and 12% of English errors.

Context-Dependent Subjectivity: Many errors occurred with sentences that appear objective in isolation but become subjective in context. The English sentence "The company's decision was announced yesterday" was classified as objective, but in context, it carried implicit criticism. This pattern accounted for 23% of English false negatives and 19% of Italian errors.

Morphological Complexity: Arabic and German frequently presented challenges due to their rich morphological structures, which often encode subjective meaning. For instance, the Arabic phrase *al-hukuma al-maz'uma* ("the so-called government") was misclassified as objective, despite the presence of the subjective marker *maz'uma* ("so-called"). Such morphologically embedded indicators of subjectivity accounted for 31% of the misclassifications in Arabic.

Cultural Context Sensitivity: Cross-cultural expressions of subjectivity posed significant challenges. Italian expressions like "come al solito" (as usual) carry implicit criticism in journalistic contexts but were often classified as objective factual statements. This pattern was particularly problematic in zero-shot transfer to Romanian, where similar Romance language expressions were misinterpreted.

6.2. Language-Specific Error Analysis

English: Despite strong overall performance, English errors concentrated in three areas: subtle sarcasm (12%), implicit bias in news reporting (20%), and ambiguous modal expressions (15%). The sentence "The politician claimed to represent the people" was classified as objective despite the potentially subjective verb "claimed."

German: German errors were dominated by compound word subjectivity (28%) and modal particle usage (22%). The compound "Scheinlösung" (apparent solution) carries inherent subjectivity that the model failed to recognize. Modal particles like "wohl" and "ja" that indicate speaker attitude were consistently misinterpreted.

Arabic: Arabic presented unique challenges with 31% of errors involving morphological subjectivity markers, 24% related to implicit cultural criticism, and 18% involving religious or political terminology with subjective connotations. The model's English-centric training poorly prepared it for Arabic's rich morphological expression of subjectivity.

Italian: Italian errors concentrated on implicit criticism (26%), conditional mood misinterpretation (21%), and cultural idiomatic expressions (18%). The subjunctive mood, which often carries subjective meaning in Italian, was frequently misclassified as objective factual reporting.

6.3. Zero-shot Transfer Error Patterns

Zero-shot transfer revealed specific patterns of cross-lingual error propagation. Romanian benefited from Italian transfer but inherited Italian-specific biases about Romanian language subjectivity markers. Polish transfer from German was partially successful but failed for culture-specific expressions.

Greek and Ukrainian errors demonstrated the limitations of script-based transfer learning. Greek errors (67% of total) involved misinterpretation of Greek-specific subjective markers that have no equivalents in Latin-script training languages. Ukrainian errors (54% of total) reflected both script challenges and cultural context misalignment.

7. Limitations and Future Work

7.1. Methodological Limitations

Our approach exhibits several fundamental limitations that constrain its effectiveness and generalizability. The reliance on English-centric pre-training creates inherent biases toward English linguistic structures and cultural contexts. This limitation is particularly evident in languages with significantly different morphological systems, such as Arabic, where the model struggles to capture language-specific subjectivity markers.

The single-model approach, while computationally efficient, limits the system's ability to capture language-specific nuances that specialized models might better handle. Unlike ensemble approaches that can leverage complementary strengths, our unified architecture must balance competing linguistic requirements across languages.

The 128-token sequence limit, chosen for computational efficiency, may truncate important contextual information necessary for accurate subjectivity detection. Many news articles contain subtle subjective cues that emerge only through extended context, which our approach cannot capture.

7.2. Data and Bias Limitations

The significant class imbalances across languages introduce systematic biases that our focal loss implementation only partially addresses. The extreme English imbalance (1:10.87 SUBJ:OBJ) may cause the model to develop conservative classification strategies that underestimate subjectivity in balanced scenarios.

Cultural bias represents a significant limitation, as our model inherits Western journalistic conventions from its English pre-training. This bias may misinterpret culturally specific expressions of subjectivity in non-Western contexts, particularly affecting Arabic performance where cultural context plays a crucial role in subjective expression.

The dataset's focus on news articles limits generalizability to other domains where subjectivity manifests differently. Social media, academic writing, or conversational contexts may require different approaches to subjectivity detection that our news-trained model cannot handle effectively.

7.3. Architectural Limitations

DeBERTa-v3-Base's architecture, while advanced, still processes text sequentially and may miss complex discourse-level patterns that span multiple sentences. Subjectivity often emerges through subtle discourse markers that require document-level understanding beyond the model's current capabilities.

The model's attention mechanism, despite its disentangled design, may not adequately capture long-range dependencies crucial for contextual subjectivity detection. Important subjective cues may be diluted across extended sequences, reducing the model's sensitivity to subtle linguistic markers.

7.4. Computational and Scalability Concerns

While our approach is more efficient than heavily multilingual models, it still requires substantial computational resources for training across multiple languages. The memory requirements for fine-tuning on large multilingual datasets may limit adoption in resource-constrained environments.

The need for language-specific fine-tuning for optimal performance constrains the model's scalability to new languages. Each additional language requires separate training and validation processes, increasing computational costs and complexity.

7.5. Ethical and Fairness Considerations

Our model may perpetuate biases present in training data, particularly regarding cultural and linguistic minorities. The English-centric pre-training may systematically underrepresent non-Western perspectives on subjectivity, potentially leading to unfair performance disparities.

The model's subjective classification decisions could impact information credibility assessments, potentially affecting how news from different linguistic communities is perceived and valued. This concern is particularly relevant for low-resource languages where model performance is degraded.

7.6. Future Research Directions

Several promising directions emerge from our analysis. Advanced multilingual pre-training approaches, such as language-adaptive pre-training or cross-lingual alignment techniques, could address the English-centric bias limitations. Incorporating explicit cultural context modeling might improve performance in non-Western languages.

Ensemble approaches combining multiple DeBERTa models with different linguistic specializations could leverage our findings while addressing single-model limitations. Hierarchical attention mechanisms that explicitly model discourse-level patterns might capture document-level subjectivity more effectively.

Data augmentation strategies, particularly for low-resource languages, could address class imbalance and cultural bias issues. Techniques such as cross-lingual data augmentation or culturally-aware synthetic data generation might improve model robustness across diverse linguistic contexts.

The integration of external knowledge sources, such as cultural context databases or linguistic resource lexicons, could provide additional signals for subjectivity detection in challenging scenarios. This hybrid approach might address some of the cultural sensitivity limitations identified in our analysis.

8. Conclusion

This work presents a comprehensive evaluation of DeBERTa-v3-Base for multilingual subjectivity detection in the context of CLEF CheckThat! 2025 Task 1. Our systematic approach across monolingual, multilingual, and zero-shot settings provides valuable insights into the capabilities and limitations of transformer-based models for cross-lingual subjectivity detection.

The results demonstrate that DeBERTa-v3-Base can achieve competitive performance across diverse linguistic contexts, with particular strength in Romance languages and reasonable zero-shot transfer capabilities. Our 6th place rankings in English and Italian monolingual tracks, combined with 5th place in Polish zero-shot transfer, indicate the viability of our approach for practical applications.

However, our analysis also reveals significant limitations, particularly in morphologically complex languages like Arabic and culturally distant languages like Greek. The performance gaps in these languages highlight the challenges of cross-lingual transfer learning and the need for more sophisticated approaches to multilingual subjectivity detection.

The focal loss implementation proved effective for handling class imbalances, though the extreme imbalances in some languages (particularly English) continue to pose challenges. Our error analysis provides detailed insights into failure modes, offering guidance for future research in this area.

The comprehensive evaluation framework we employed—spanning monolingual, multilingual, and zero-shot settings—provides a template for future work in multilingual subjectivity detection. The systematic error analysis and limitation discussion offer practical insights for researchers working on similar tasks.

Future work should focus on addressing the cultural bias limitations identified in our analysis, developing more sophisticated cross-lingual alignment techniques, and exploring hybrid approaches that combine the efficiency of our method with the performance advantages of ensemble systems. The integration of cultural context modeling and advanced multilingual pre-training represents particularly promising directions for advancing the field.

Declaration on Generative AI

We affirm that all experimental design, implementation, training, evaluation, and analysis were conducted manually by the authors. Generative AI tools such as ChatGPT or similar were not used for writing technical content, generating code, or automating evaluations. We used Grammarly exclusively for minor language corrections (grammar and punctuation). No parts of the scientific contributions were generated using AI writing tools. The integrity and originality of the work have been maintained in accordance with the CLEF 2025 submission guidelines.

References

[1] J. Wiebe, R. Bruce, T. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 246–253. URL: https://aclanthology.org/P99-1032/. doi:10.3115/1034678.1034733.

[2] CLEF Organizers, Checkthat! lab at clef 2025, 2025. URL: https://checkthat.gitlab.io/clef2025/, accessed: 2025-06-30.

[3] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021). URL: https://arxiv.org/abs/2111.09543.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[6] E. Riloff, J. Wiebe, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 25–32. URL: https://aclanthology.org/W03-0404/. doi:10.3115/1119176.1119180.

[7] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432. URL: https://aclanthology.org/D15-1167/. doi:10.18653/v1/D15-1167.

[8] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181/. doi:10.3115/v1/D14-1181.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2020). URL: https://arxiv.org/abs/1911.02116.

[10] J. Barnes, R. Klinger, S. S. i. Walde, Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 818–829. URL: https://aclanthology.org/P19-1078/. doi:10.18653/v1/P19-1078.

[11] CLEF Organizers, Checkthat! lab at clef 2023, 2023. URL: https://ceur-ws.org/Vol-3492/, cEUR Workshop Proceedings.

[12] K. Schlicht, et al., Dwreco at checkthat! 2023: Gpt-3 based data augmentation for subjectivity detection, in: CLEF 2023 Working Notes, volume 3492 of *CEUR Workshop Proceedings*, 2023, pp. 1–12. URL: https://arxiv.org/abs/2309.06846. arXiv:2309.06846, also available at CEUR-WS Vol. 3492.

[13] G. Pachov, I. Koychev, Gpachov at checkthat! 2023: Ensemble models for multilingual subjectivity detection, in: CLEF 2023 Working Notes, volume 3492 of *CEUR Workshop Proceedings*, 2023, pp.

1–12. URL: https://arxiv.org/abs/2309.06844. arXiv:2309.06844, also published in CEUR-WS Vol. 3492.

[14] E. Casanova, L. da Silva, Hybrinfox at checkthat! 2024: Hybrid roberta-vago for multilingual subjectivity, in: CLEF 2024 Working Notes, volume 3660 of *CEUR Workshop Proceedings*, 2024, pp. 1–12. URL: https://arxiv.org/abs/2406.07160. arXiv:2406.07160, also published in CEUR-WS Vol. 3660.

[15] A. Singh, N. Sharma, Multilingual transfer for subjectivity detection: Nullpointer at checkthat! 2024, in: CLEF 2024 Working Notes, volume 3660 of *CEUR Workshop Proceedings*, 2024, pp. 1–12. URL: https://arxiv.org/abs/2406.07159. arXiv:2406.07159, also published in CEUR-WS Vol. 3660.

[16] T. Leistra, Subjectivity Classification Using DeBERTa and its Multilingual Variants, Master's thesis, Utrecht University, 2023. URL: https://studenttheses.uu.nl/handle/20.500.12932/45452.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 318–327. URL: https://arxiv.org/abs/1708.02002. doi:10.1109/TPAMI.2018.2858826.

[18] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018, pp. 66–71. URL: https://aclanthology.org/D18-2012/. doi:10.18653/v1/D18-2012.