Fine-tuning of mDeBERTaV3 & ModernBERT for Subjectivity Detection
3-cfu Project Work

Matteo Fasulo, Luca Babboni and **Luca Tedeschini**
Master's Degree in Artificial Intelligence, University of Bologna
{ matteo.fasulo, luca.babboni2, luca.tedeschini3 }@studio.unibo.it

Abstract

Detecting subjectivity in natural language is crucial for various NLP tasks, including fake news detection, fact-checking and automatic summarization of reviews. However, achieving robust subjectivity detection across different languages remains a challenging task due to the complexity of linguistic diversity and cultural differences. In this study, we present our system developed for 2025 CheckThat! Lab task 1 on subjectivity detection. We employ two distinct approaches: one using BERT-like architectures with *mDeBERTaV3-base* [1] and *ModernBERT-base* [2], and the other using LLMs such as *Llama3.2-1B* [3]. The results indicate that BERT-like models exhibit superior performance in capturing nuanced information and accurately determining subjectivity compared to LLMs. Moreover, we find that incorporating sentiment information yields significant improvements in the subjective F1 score for English and Italian languages, whereas the improvement is marginal for the others. This means that subjectivity is identified better when the sentiment information is injected into the model. Furthermore, to overcome the imbalanced label distributions, we employed a decision threshold calibration procedure. This procedure has a substantial impact on performance for languages with imbalanced label distributions, while providing only marginal gains for more balanced languages.

1 Introduction

The rapid proliferation of online content has generated an immense volume of unstructured data, complicating the differentiation between subjective opinions and objective facts. This distinction is essential for a variety of applications, including misinformation detection and automated fact-checking.

As news production and public discourse become increasingly polarized, the accurate classification of subjective and objective statements has become progressively important: it is essential to be able to distinguish objectivity when reading news paper.

Subjectivity detection constitutes a well-established problem within the domain of NLP, commonly considered a subtask of sentiment analysis. Traditional methodologies have encompassed rule-based frameworks, lexicon-based techniques, and classical machine learning models employing hand-crafted features (Kamal, 2013). Although these methods demonstrate satisfactory performance in controlled environments, they often encounter challenges related to linguistic variability, contextual dependencies, and cross-linguistic generalization. Recent advances in deep learning, particularly transformer-based language models, have markedly improved classification performance (Savinova and Moscoso Del Prado, 2023). Nonetheless, many pre-trained models are tailored for monolingual tasks, which constrains their efficacy in multilingual settings.

This study aims to address these limitations by developing a subjectivity detection system for the CLEF 2025 CheckThat! Lab Task 1. The task requires the classification of sentences extracted from news articles in multiple languages, namely Arabic, German, English, Italian, and Bulgarian, as either subjective (SUBJ) or objective (OBJ). The classification is carried out at the sentence level without access to the surrounding context, thereby increasing the complexity of the task. To address this challenge, we employ two distinct approaches: one based on BERT-like architectures and the other on LLMs.

BERT-like architectures are advantageous in their ability to generate contextualized representations that enhance the understanding of whether a sentence conveys objective content. Notably, previous editions of the CLEF CheckThat! Lab

[1] microsoft/mdeberta-v3-base
[2] answerdotai/ModernBERT-base
[3] meta-llama/Llama-3.2-1B

have have been tackled using such models to detect subjectivity within news articles (Leistra and Caselli, 2023). In our approach, we evaluate the mDeBERTaV3-base (He et al., 2021b,a), a multilingual transformer-based renowned for its robust cross-linguistic generalization capabilities, alongside the ModernBERT-base (Warner et al., 2024), a recently developed English-only model designed to achieve efficiency and performance improvements with fewer parameters. To further enhance subjectivity detection, we experimented with data augmentation techniques by integrating sentiment values associated with each sentence.

The second approach explores LLM fine-tuning using Llama-3.2-1B with an added classification head to specialize the model for subjectivity detection using English-only data.

Our experimental framework involves fine-tuning these models on language-specific datasets provided by the challenge organizers [4] as well as a combined dataset from all languages to examine cross-lingual capabilities. Performance is evaluated using metrics such as macro-average F1 score and positive class (SUBJ) F1 score following a decision threshold calibration procedure.

The empirical results indicate that BERT-like models demonstrate a superior ability to capture nuanced information and accurately determine subjectivity compared to LLMs. Furthermore, augmenting sentences with sentiment information resulted in substantial performance improvements. Additionally, the decision threshold calibration procedure significantly enhanced performance for the most unbalanced languages, such as Italian and Arabic, while its impact was notably smaller for nearly balanced languages, including Bulgarian, English, and German. This finding is consistent with the results presented in (Abdelhamid and Desai, 2024), which highlight the effectiveness of threshold calibration in addressing class imbalance.

2 Background

The CLEF 2025 CheckThat! Lab Task 1 comprises three distinct subtasks, each corresponding to a different experimental setting. Monolingual, where both training and testing are conducted in a single language; multilingual, where training and testing involve multiple languages; and zero-shot, where the model is trained on several languages but evaluated on previously unseen languages.

[4] CLEF 2025 CheckThat! Lab Task 1 Data

The multilingual and zero-shot subtasks impose additional constraints, as they require the use of multilingual models to correctly process tokenized inputs. During our work with the multilingual setting, we observed notable variations in model performance across different languages. Certain fine-tuning parameters and procedures performed well on some languages but significantly worse on others. We hypothesize that these discrepancies stem from intrinsic linguistic differences; however, due to our limited expertise in many of the languages included in the task, we were unable to investigate this hypothesis in depth.

About the architectural differences between mDeBERTa-V3 and ModernBERT, mDeBERTa-V3 employs a *disentangled attention mechanism* that separately processes content and positional information, enabling nuanced contextual analysis through gradient-disentangled embedding sharing. In contrast, ModernBERT adopts *rotary positional embeddings (RoPE)* and a hybrid attention system alternating between global context modeling (8,192-token capacity) and localized 128-token windows for computational efficiency. mDeBERTa-V3 utilizes a 12-layer transformer with 86M backbone parameters, optimized for classification tasks through enhanced mask decoding. ModernBERT implements a deeper 22-layer base architecture (149M parameters) with pre-normalization layers and GeGLU activations, specifically engineered for long-sequence processing and code comprehension through strategic sequence packing and unpadding techniques.

Given the aforementioned reasons, we selected mDeBERTa due to its result in NLU tasks [5] and ModernBERT due to comparable performances with mDeBERTa while being more computationally efficient. [6]

3 System description

The implemented system consists of a straightforward pipeline designed to process data, train models, and generate predictions. The architecture and coding aspects of the system are described in the following.

The base architecture follows a standard fine-tuning and evaluation pipeline leveraging pre-trained language models. The architectural design for the classification heads follows a simple feed-forward

[5] mDeBERTa-v3-base fine-tuning on NLU tasks

[6] ModernBERT-base model comparison

neural network having as input the `CLS` token embedding. As for the Llama3.2-1B based architecture, the classification head has its input dimension coherent with the *LLM* output. The sentiment architecture, only employed on the mDeBERTa model, takes into account sentiment by having a larger input size in its classification head, to accommodate the new features that are concatenated to the last hidden state of the model.

Data Preparation

For the sentiment pipeline, as visible in Figure 1, we augmented each dataset with the output provided by the *twitter-xlm-roberta-base-sentiment* (Barbieri et al., 2022) model, which is a multilingual XLM-roBERTa-base model finetuned for sentiment analysis. The output produced by the model is a three dimensional vector representing the *positive*, *neutral*, and *negative* score of the sentiment of a sentence.

Tokenization

For each dataset, we used the tokenizer specific to the model being fine-tuned. During tokenization, sentences were padded and truncated to a maximum length of 256 tokens. This approach successfully encompassed over 75% of sentence lengths across all languages in the dataset.

Training and Inference

The models themselves are fine-tuned on the *train* dataset, hyperparameter tuning is instead performed on the *dev* dataset and *dev-test* dataset is used to test the model on unseen data.

Post-Processing

To address class imbalance and enhance the balance between precision and recall, we transformed the raw logits into probability scores using the *Softmax* function. This transformation provides two complementary probabilities for class membership: OBJ (0) and SUBJ (1). To derive the final predictions, we implemented a threshold optimization process. This involved conducting a grid search over a specified interval to identify the optimal decision threshold for the positive class (SUBJ). The grid search was performed over the interval $(0.1, 0.9)$ with 100 iterations, aiming to maximize the macro F1 score on the validation set. Once the optimal threshold was determined, it was applied to the test set to generate the final predictions.

The data preparation, tokenization, threshold optimization, and prediction generation modules were entirely written by us using the *Huggingface* libraries. Model training and evaluation pipelines were built upon existing Huggingface training utilities but were heavily modified to incorporate customized evaluation metrics and threshold tuning. The base architectures for mDeBERTaV3-base, ModernBERT-base, and Llama3.2-1B were obtained from the Huggingface model hub without modification, aside from the addition of a classification head for Llama3.2-1B, while the sentiment architecture is based upon the mDeBERTaV3-base with its last classification head modified to accomodate the additional sentiment features.

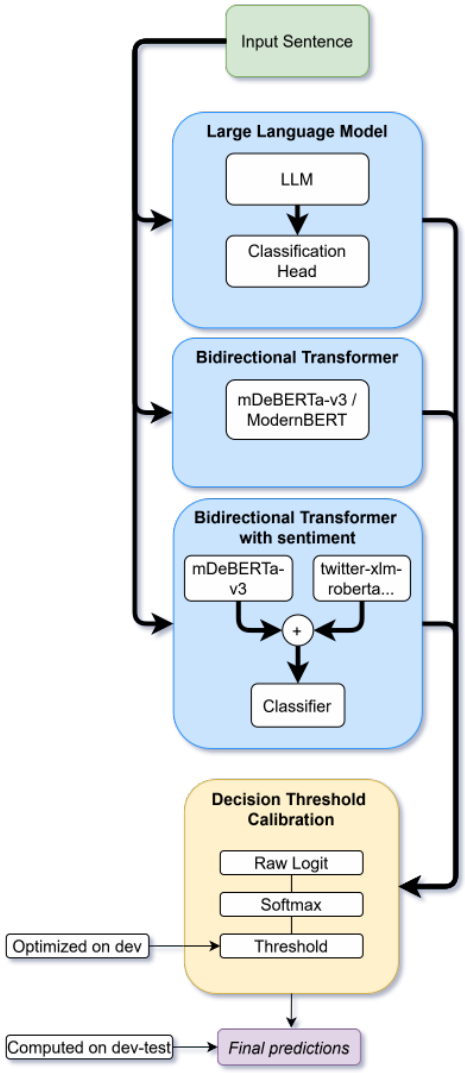

Figure 1: Model architecture with the three developed pipelines.

4 Data

The dataset provided for the challenge is divided

into training, validation, and test sets. The ground truth for the sentence is annotated by the dataset creator according to the subjectivity detection guidelines outlined in (Antici et al., 2024) which are the following: sentences are classified as subjective if they express personal opinions, sarcasm, exhortations, discriminatory language, or rhetorical figures conveying an opinion. Objective sentences include third-party opinions, open-ended comments, and factual conclusions. Additionally, reported speech is treated as objective, and statements reflecting personal emotions without forming an opinion are also classified as objective to maintain real-world applicability. Table 1 presents the distribution of labels across the training, validation, and test sets. As a first observation, we can see that the label distribution is unbalanced, with the *objective* tag being the most prominent across all languages. A closer inspection reveals that Italian is the most unbalanced language, exhibiting an objective-to-subjective ratio of approximately 4:1, followed by Arabic. To address this issue, we applied various techniques during the design of our solution. Among them, thresholding optimization yielded the most promising results.

Language	Training		Dev		Dev-Test	
	OBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ
Arabic	1,391	1,055	266	201	425	323
Bulgarian	406	323	175	139	143	107
English	532	298	240	222	362	122
German	492	308	317	174	226	111
Italian	1,231	382	490	177	377	136

Table 1: Distribution of objective (OBJ) and subjective (SUBJ) labels across different languages and dataset splits. The table presents statistics for the training, development (Dev), and development-test (Dev-Test) sets.

5 Experimental setup and results

5.1 Task 1 - Monolingual

For the monolingual task, we fine-tuned three pretrained models: mDeBERTaV3-base, ModernBERT-base, and Llama3.2-1B. Due to hardware limitations, we quantized Llama3.2-1B to 8-bit using Low-Rank Adaptation (LoRA). Each model was fine-tuned and evaluated exclusively on a specific language to maintain consistency. Our initial results indicated that simple fine-tuning yielded satisfactory outcomes for all languages except Arabic.

To address this issue, we applied a pre-translation technique to the Arabic dataset by translating it into English—a common approach in NLP tasks. The translation was performed using the *Helsinki-NLP/opus-mt-ar-en* (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020) model developed by the Language Technology Research Group [7] at the University of Helsinki.

The finetuning process was performed for 6 epochs with a batch size of 16 and a learning rate of 1×10^{-5}. AdamW optimizer with a linear scheduler and warmup was used while having Cross Entropy Loss with class weights to address class imbalance. For each language, the model with the lowest validation loss during finetuning was selected.

We assessed model performance using two evaluation metrics. The *Macro Average F1 Score* measures overall performance across both classes (OBJ and SUBJ), treating each class equally regardless of its frequency. Additionally, the *Positive Class (SUBJ) F1 Score* specifically evaluates the effectiveness of subjectivity detection by focusing on the positive class (SUBJ). This metric is particularly insightful, as our models have struggled to correctly identify subjective sentences. By analyzing this score, we can determine improvements specifically related to the subjective class.

Language	Macro F1	SUBJ F1
mDeBERTa-V3		
Arabic	0.5805	0.5598
Bulgarian	0.7555	0.7222
English	0.6375	0.4046
German	0.8218	0.7652
Italian	0.7654	0.6291
mDeBERTa-V3 + Sentiment		
Arabic	0.5735	0.5741
Bulgarian	0.7718	0.7407
English	0.7036	0.5279
German	0.8291	0.7759
Italian	0.7769	0.6804
ModernBERT		
English	0.6922	0.5612
Llama3.2-1B		
English	0.6375	0.4046

Table 2: Model performance for Task 1 across languages. The table reports Macro F1 and Positive Class (SUBJ) F1 scores using a decision threshold calibration procedure.

[7] helsinki language-technology

Language	Threshold		No Threshold	
	Macro F1	SUBJ F1	Macro F1	SUBJ F1
Arabic	**0.5805**	**0.5598**	0.5538	0.4184
Bulgarian	**0.7555**	**0.7222**	0.7491	0.6970
English	0.6375	0.4046	**0.7069**	**0.5556**
German	**0.8218**	0.7652	0.8217	**0.7699**
Italian	**0.7654**	**0.6291**	0.7048	0.6237

Table 3: Comparison of model performance with and without thresholding across different languages. The table reports Macro F1 and Subjective (SUBJ) F1 scores for both approaches.

5.2 Task 2 - Multilingual

For the multilingual task, we fine-tuned the mDeBERTaV3-base model on a combined dataset encompassing all language-specific datasets. This approach was chosen because ModernBERT is monolingual, and training Llama3.2-1B on the entire dataset was computationally prohibitive due to resource constraints. We then evaluated the model's performance across all languages to assess its generalization capabilities.

The training and evaluation procedures followed the same methodology as in Task 1.

Language	Macro F1	SUBJ F1
mDeBERTa-V3		
Multilingual	0.6942	0.6114
Excluding Arabic	0.7817	0.6887
mDeBERTa-V3 + Sentiment		
Multilingual	0.6798	0.5332
Excluding Arabic	0.7962	0.7114

Table 4: Evaluation results of mDeBERTa-V3 on multilingual data.

Given the known challenges associated with the Arabic language, we conducted experiments by excluding Arabic from the dataset. As anticipated, this approach resulted in a notable improvement, with the macro F1-score increasing by approximately 0.1 and the subjective F1-score by ~ 0.2.

5.3 Task 3 - Zero-shot

For the zero-shot task, we constructed ad hoc language-specific training datasets containing three languages and test language-specific datasets containing the remaining two, evaluating all possible combinations.

Since presenting all results would take up too much space (a total of 18 rows), we decided to report only some experimental results in Table 5.

Language	Model	Macro F1	SUBJ F1
Ar, Bg, Ge	Base	0.7395	0.6066
Ar, Bg, Ge	Base + Sentiment	0.7461	0.6134
En, It	Base	0.6147	0.5166
En, It	Base + Sentiment	0.6121	0.5087

Table 5: Zero Shot performances across different language-specific dataset. If not specified in the language column, the language-specific dataset is used to test the model.

5.4 Analysis of the Sentiment Pipeline Benefits

The above results indicate that incorporating the sentiment into models tends to improve the SUBJ F1 score across nearly all tested languages. In particular, for English, the addition of sentiment information yields considerable improvement. With the baseline mDeBERTa-V3, the English SUBJ F1 score is 0.4046. After adding the sentiment pipeline (mDeBERTa-V3 + Sentiment), the SUBJ F1 score increases to 0.5279—a gain of over 30% relative to the baseline. Additionally, the overall Macro F1 also increases from 0.6375 to 0.7036, demonstrating that sentiment-based features contribute positively not only to the positive class but also to the general performance.

To further understand these improvements, we conducted an in-depth analysis on the English test set. In Tables 6 and 7, we examine the sentiment scores associated with correctly and incorrectly classified sentences, respectively. This analysis provides insights into how sentiment processing helps adjust model decisions in favor of more accurate subjectivity predictions. As discussed further in Section 6, these sentiment values offer cues that refine the decision threshold, especially in ambiguous cases, leading to the improvements seen in the SUBJ F1 scores.

Label	Mean			Std		
	Positive	Neutral	Negative	Positive	Neutral	Negative
OBJ	0.32	0.31	0.36	0.20	0.31	0.32
SUBJ	0.23	0.24	**0.51**	0.19	0.35	0.35

Table 6: Mean and standard deviation of sentiment values when the sentiment model correctly identifies sentences, but the other model fails.

6 Discussion

6.1 Quantitative Results

The results from our experiments reveal several key insights into the performance of different models

Label	Mean			Std		
OBJ	Positive	Neutral	Negative	Positive	Neutral	Negative
	0.23	0.37	0.39	0.14	0.32	0.40
SUBJ	0.29	0.37	0.32	0.23	0.36	0.34

Table 7: Mean and standard deviation of sentiment values when the sentiment model does not correctly identifies sentences, but the other model does.

and configurations for subjectivity detection across multiple languages.

The Llama3.2-1B model did not perform as well as the BERT-like models. This suggests that while LLMs have potential, they may require more specialized fine-tuning or architectural adjustments to match the performance of BERT-like models in subjectivity detection tasks. Additionally, larger models without weight quantization might achieve better performance; however, such models were not feasible given our resource constraints.

Although our primary goal was to improve predictions for the Arabic language, the translation process may have introduced noise and inaccuracies. Indeed, the pre-translation technique resulted in poorer performance compared to the non-translated approach. This additional noise can propagate through the model, thereby diminishing its classification power.

The integration of sentiment information has notably enhanced the Positive Class (SUBJ) F1 score for languages such as Italian and English. This improvement indicates that sentiment features can be especially advantageous for languages where identifying subjectivity presents more challenges, potentially due to linguistic subtleties or the typical expression of subjective content in these languages. The observed gains in performance suggest that sentiment analysis helps refine the model's ability to discern subjective nuances, thereby improving classification accuracy. However, the precise mechanisms driving this enhancement are not fully understood and merit further exploration to uncover the underlying factors contributing to these improvements.

Secondly, the decision threshold calibration had a significant impact on performance for languages with highly imbalanced datasets, such as Arabic and Italian. For these languages, the decision threshold adjustment led to substantial improvements in both Macro F1 and SUBJ F1 scores. In contrast, for more balanced languages like Bulgarian, English, and German, the impact of threshold calibration was less pronounced. This indicates that threshold calibration is a crucial step for handling class imbalance effectively.

The Arabic monolingual model exhibits the weakest performance among the models analyzed, which may be attributed to multiple factors, including limitations in the pretraining data, domain mismatches, or structural differences in how information is encoded in Arabic compared to other languages. These challenges suggest that the model's training data may lack sufficient diversity or coverage, or that inherent linguistic characteristics of Arabic present difficulties for the model's learning process. To further illustrate these weaknesses, we observe that when Arabic is included in the zero-shot task test dataset, the performance drops significantly compared to when the language is present in the training dataset. This suggests that even a multilingual training approach is not sufficient to address these challenges. When removing Arabic from the dataset in both the multilingual and zero-shot subtasks, we observed a significant improvement in performance, further proving the language specificity. While further investigation is required to pinpoint the exact cause of this performance gap, such an analysis falls outside the scope of this study.

In the multilingual task, the inclusion of sentiment information led to improved performance compared to using the mDeBERTa-V3 model alone. This suggests that sentiment information can enhance the model's ability to generalize across languages, likely by providing additional context that aids in distinguishing subjective content.

To better understand the impact of sentiment information on the performance of our models, we conducted a cross-error analysis on the sentences misclassified by both the sentiment model and the baseline model, as reported in Section 5. Additionally, we examined the distribution of sentiment across the dataset. In our initial analysis, we identified the sentences correctly classified by the sentiment model but misclassified by the baseline model (Table 6). The results indicate that the average negative sentiment for subjective sentences is significantly higher compared to the baseline (0.33). This suggests that the sentiment model is more effective in recognizing subjectivity when the sentiment is negative. To validate this intuition, we performed the same analysis on the sentences misclassified by the sentiment model but correctly classified by the baseline model (Table 7). In this case, no signifi-

cant pattern emerged, as all metrics remained close to their baseline values. Finally, we analyzed the overall sentiment distribution in the dataset (Figure 2). The results confirm our initial observations: subjective sentences tend to exhibit a more negative sentiment on average, reinforcing the identified pattern by our model.

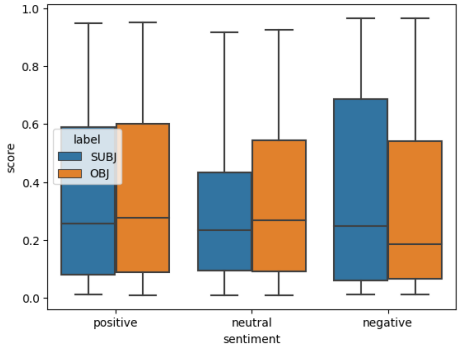

Figure 2: Sentiment distribution over the dataset. The black line represents the median of the distribution, the box represents the interquartile range while the whiskers represents the variability of the data (minimum and maximum value)

6.2 Error Analysis

Since we discovered interesting results by incorporating sentiment into the classification process, our error analysis focuses on comparing misclassifications between the base model and the sentiment model. By understanding the improvements and identifying sentences that remain misclassified, further studies can build upon our findings as a starting point. Starting with English, we plotted the violin plot by sentiment and label (Figure 3). Observing this distribution, we confirm our previous findings: most subjective sentences exhibit a very high negative sentiment score. Specifically, the median is above 0.6, while the third quartile exceeds 0.8. This trend is also observed in the Italian language. Conversely, Arabic and Bulgarian exhibit the opposite pattern, as shown in Figure 4. Given that Arabic is the most prominent language in the combined datasets (Table 1), these adverse trends likely hindered the model's ability to extract sentiment information from the dataset. Indeed, when Arabic is removed from the multilingual datasets, performance improves.

To better illustrate the advantages of incorporating sentiment into the classification process, we present two examples of subjective sentences that were correctly classified by the sentiment model but misclassified by the base model. For each sentence, we report the corresponding sentiment score in the format (positive, neutral, negative).

But then Trump came to power and sidelined the defense hawks, ushering in a dramatic shift in Republican sentiment toward America's allies and adversaries. $(0.109, 0.035, \mathbf{0.856})$

Boxing Day ambush & flagship attack Putin has long tried to downplay the true losses his army has faced in the Black Sea. $(0.056, 0.014, \mathbf{0.930})$

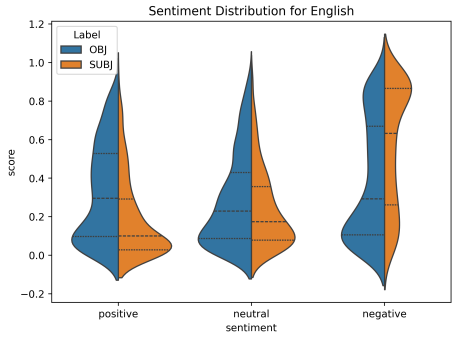

Figure 3: Sentiment distribution over the english language. The three lines in the violin plot represents the first, second and third quartile.)

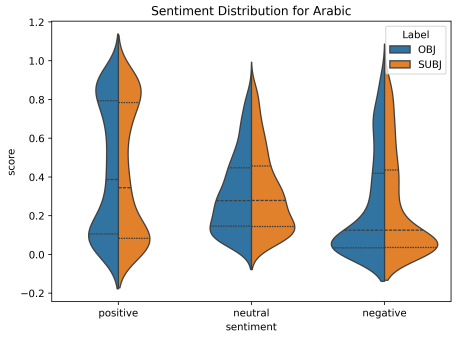

Figure 4: Sentiment distribution over the Arabic language. The three lines in the violin plot represents the first, second and third quartile.)

7 Conclusion

In this study, we explored the effectiveness of different fine-tuned transformer-based models for subjectivity detection across multiple languages. Our findings demonstrate that BERT-like architectures outperform LLMs in capturing the nuanced information necessary for distinguishing between subjective and objective content. Specifically, mDeBERTaV3-base with sentiment augmentation achieved the highest performance across

most languages, with notable improvements in English and Italian. The decision threshold calibration procedure proved particularly effective for languages with imbalanced class distributions, significantly enhancing both macro-average F1 and subjective class F1 scores. Interestingly, we observed consistent challenges with Arabic language detection across all experimental settings, suggesting potential limitations in how current multilingual models encode information for this language. Our error analysis revealed that complex linguistic structures and context-dependent expressions represent persistent challenges for all models tested. Future work could explore more sophisticated approaches for feature fusing (rather than simple concatenation), additional linguistic features beyond sentiment, and investigate specialized approaches for languages with unique structural characteristics. Overall, our study demonstrates the effectiveness of BERT-like models for subjectivity detection and highlights the importance of considering linguistic variability, contextual dependencies, and class imbalance in multilingual settings. Further research is needed to explore the potential of LLMs and to address the challenges identified in our error analysis.

8 Links to external resources

- Github repository
- Dataset

References

Mohamed Abdelhamid and Abhyuday Desai. 2024. Balancing the scales: A comprehensive study on tackling class imbalance in binary classification.

Francesco Antici, Federico Ruggeri, Andrea Galassi, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2024. A corpus for sentence-level subjectivity detection on English news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 273–285, Torino, Italia. ELRA and ICCL.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ahmad Kamal. 2013. Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources.

Folkert Atze Leistra and Tommaso Caselli. 2023. Thesis titan at checkthat! 2023: Language-specific fine-tuning of mdebertav3 for subjectivity detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, CEUR Workshop Proceedings, pages 351–359. CEUR Workshop Proceedings (CEUR-WS.org). Publisher Copyright: © 2023 Copyright for this paper by its authors.; 24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023 ; Conference date: 18-09-2023 Through 21-09-2023.

Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314, Toronto, Canada. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.