Subjectivity in News Articles
Project Work

Mehregan Nazarmohsenifakori, Shafagh Rastegari, Mohammad Pourtaheri
Master's Degree in Artificial Intelligence, University of Bologna
{ mehrega.nazarmohseni, shafagh.rastegari, mohammad.pourtaheri }@studio.unibo.it

Abstract

The ability to detect subjectivity in natural language is critical for numerous applications, including sentiment analysis, argument mining, and fake news detection. Nevertheless, achieving effective and accurate subjectivity detection across diverse languages poses significant challenges, primarily due to inherent linguistic variations. This report details the system developed for the 2025 CheckThat! Lab Task 1, focusing on subjectivity detection in news article. Our approach leverages the multilingual mDeBERTaV3-base model. Specifically, we fine-tune mDeBERTaV3-base models for each language. To specialize these systems for target languages and mitigate the impact of class imbalance within the training data, we incorporate language-specific development data during the fine-tuning process.

1 Introduction

Subjectivity detection acts as a foundational step for sentiment analysis, argument mining, and fake news detection, helping to differentiate between verifiable facts and personal opinions or biases. However, the task of effectively and accurately detecting subjectivity becomes considerably more complex when extended across diverse languages. Each language possesses unique linguistic structures, idiomatic expressions, and cultural nuances that can significantly influence how subjectivity is conveyed.

This report details the system developed to address these challenges as part of the 2025 CheckThat! Lab Task 1 (in News Articles, 2025), which specifically focuses on subjectivity detection within news articles, provides an multilingual setting (Arabic, Greek, Bulgarian, Ukrainian, Romanian, English, German, Italian, and Polish) for the identification of the subjectivity status of sentences. The task is framed as a binary classification, whose goal is to assign the subjectivity status (SUBJ for subjective, and OBJ for objective) to a given sentence. Our proposed approach centers on the utilization of the multilingual mDeBERTaV3-base model (He et al., 2023), a powerful pre-trained transformer architecture known for its strong performance across various natural language processing tasks.

To adapt this robust model to the specific requirements of multilingual subjectivity detection, we employ a fine-tuning strategy. We utilize the provided datasets to our models for each target language and evaluate with Precision, Recall, and F1 of the SUBJ class and the macro-averaged scores. Then compared our result to the baseline model which provided in the description of the Lab page. The baseline model is a logistic regressor trained on a Sentence-BERT multilingual to classify the OBJ and SUBJ. Section 2 describes the data used in this project. Section 3 presents the model for each experimental setting. Section 4 details the hyperparameters, evaluation metrics, and language-specific configurations for each model. Section 5 reports the results for each language and setting, and Section 6 compares these results with the baseline model.

2 Data

The data in this project is provided in the Lab page. It contains train, test, and validation dataset for five languages: English, Italian, German, Bulgarian, and Arabic. For the other languages: Ukrainian, Polish, Greek, and Romanian, there is no training dataset, so we used them to test our zero-shot setting.

Table 1 provides an overview of the label distributions for the languages included in this project. The data is provided in TSV format, containing *sentence_id*, *sentence*, and *label* fields. Additionally, unlabeled test datasets are available for each language.

Language	Train		Validation	
	SUBJ	OBJ	SUBJ	OBJ
English	298	532	240	222
Italian	382	1231	177	490
German	308	492	174	317
Bulgarian	323	406	139	175
Arabic	1055	1391	201	266

Table 1: Data distribution per language.

For multilingual analyzes, the training set was not provided and was constructed using two distinct approaches, each of which will be evaluated for its results. The first approach involved combining the training sets of five languages: English, German, Italian, Arabic, and Bulgarian. The second approach involved creating a balanced dataset from these same five languages, which also contained a 60% OBJ and 40% SUBJ distribution.

3 System description

The task is structured into three distinct settings, and this section details the proposed approach for each. For the majority of these settings, the **mdeberta-v3-base** model was fine-tuned (Leistra and Caselli, 2023). This model is the multilingual variant of the original DeBERTa architecture (He et al., 2021). However, a different model used for some languages in monolingual which specified within its dedicated section.

3.1 Monolingual

For this part, we train the model on each 5 languages: English, German, Italian, Arabic, and Bulgarian. Then test each resulted model on their test datasets. For German, Italian, and Arabic we also experiment on the models which pre-trained on these languages. For German used **German Bert** (team , dbmdz) which is the model based on Bert for this language. For Italian, use **UmBERTo** (Parisi et al., 2020) which is a Roberta-based Language Model trained on large Italian Corpora. For Arabic, use **AraELECTRA** (Antoun et al., 2021).

3.2 Multilingual

Here we fine-tune with two distinct datasets, combined dataset and balanced dataset, which was proposed in Section 2. Then test on the test data provided for the multilingual setting, after that test the best model on each test data for each of the languages.

3.3 Zero-shot

Here we use the best model from the multilingual part and test on the unseen languages: Greek, Polish, Ukrainian, and Romanian.

4 Experimental setup

Fine-tuning of the pre-trained models was conducted with specific hyperparameters, which were fit for each language and models which described in Section 3. The detailed hyperparameters for each language and setting are presented in Table 2.

4.1 Evaluation Metrics

Model performance was primarily assessed using the Macro F1 score. This metric is particularly robust for classification tasks, especially those affected by class imbalance, as it computes the F1 score for each class independently and then averages them. In addition to Macro F1, Precision, Recall, and the F1 score specifically for the "SUBJ" (subjective) class were reported.

We calculated all these metrics in three settings and compared them in the following section.

5 Discussion

This section presents the experimental outcomes across monolingual and multilingual, and zero-shot settings by assessing model performance.

5.1 Monolingual Performance Analysis

This analysis compares the performance of the mDeBERTaV3-base model and language-specific models (German BERT, UmBERTO, AraELECTRA) across English, German, Italian, Arabic, and Bulgarian datasets in a monolingual setting. The results are summarized in Table 3.

Languages	Macro F1	Precision	Recall	SUBJ F1
English	0.71735	0.57447	0.57447	0.60335
Germany	0.75289	0.68750	0.65254	0.66957
Germany(BERT base)	0.78347	0.74074	0.74074	0.70796
Italian	0.77075	0.68696	0.73832	0.71171
Italian(UmBERTo)	0.73341	0.67677	0.62617	0.65049
Arabic	0.57380	0.39377	0.44984	0.41994
Arabic(AraELECTRA)	0.59194	0.40650	0.64725	0.49938
Bulgarian	0.73918	0.69725	0.71028	0.70370

Table 3: Result of the Monolingual setting. We report macro F1, Precision, and Recall, as well as the F1 for the SUBJ class.

In German, the German (BERT base) model demonstrated a significant performance advantage over mDeBERTaV3-base, achieving a Macro F1 of 0.78347 compared to mDeBERTaV3-base's

Languages	Batch Size	Epoch	LR	Warmup Steps	Warmup Ratio	WeightDecay
English	32	6	3e-5	6	_	0.01
Germany	16	6	2e-5	_	0.1	0.1
Germany(BERT base)	16	6	2e-5	_	0.1	0.1
Italian	32	6	5e-5	_	0.15	0.1
Italian(UmBERTo)	32	6	5e-5	_	0.15	0.1
Arabic	16	3	4e-5	_	0.08	0.2
Arabic(AraELECTRA)	16	3	6e-5	_	0.4	0.3
Bulgarian	16	6	2e-5	_	0.1	0.1
Multilingual	32	4	5e-5	500	_	0.3
Multilingual Balanced	32	6	5e-5	500	_	0.3

Table 2: (Hyper)parameters that were using form DeBERTaV3 per language and Multilingual setting and three additional models for Germany, Italian and Arabic.

0.75289. This improvement also extended to the SUBJ F1 score (0.70796 versus 0.66957), clearly indicating the benefit of utilizing a model pre-trained specifically on German data for this task.

Conversely, in Italian, mDeBERTaV3-base exhibited better performance compared to the language-specific Italian (UmBERTO) model, with a Macro F1 of 0.77075 against UmBERTO's 0.73341. This finding suggests that for Italian, the broader pre-training of the multilingual mDeBERTaV3-base model was more effective for subjectivity detection than the Roberta-based UmBERTO.

For Arabic, both mDeBERTaV3-base (Macro F1: 0.57380) and Arabic (AraELECTRA) (Macro F1: 0.59194) yielded the lowest performance among all languages evaluated in the monolingual setting, therefore, Arabic proved to be the most challenging language for subjectivity detection.

5.2 Multilingual Performance Analysis

This section compares the performance of two distinct multilingual training strategies: one utilizing a simply combined dataset and another employing a specifically balanced dataset. The results are presented in Table 4.

Languages	Macro F1	Precision	Recall	SUBJ F1
Multilingual	0.68308	0.50656	0.81099	0.62360
Multilingual Balanced	0.72472	0.61240	0.63813	0.62500

Table 4: Results for the Multilingual setting

There is a substantial improvement in Macro F1 and Precision for the balanced approach. This approach, by providing a more equitable representation of classes during training, enables the model to learn more discriminative features, resulting in a more reliable and accurate classifier. This highlights that data preparation, particularly balancing, is critical for our architecture.

The Table 5 show how effectively the best multilingual model generalizes when tested on individual language test datasets. It also demonstrate strong cross-lingual transfer capabilities.

Languages	Macro F1	Precision	Recall	SUBJ F1
English	0.69961	0.53333	0.65882	0.58947
Italian	0.78619	0.67969	0.81308	0.74043
Arabic	0.68439	0.54969	0.57282	0.56101
Bulgarian	0.72036	0.71739	0.61682	0.66332
Germany	0.78946	0.79787	0.63559	0.70755

Table 5: Results for the Multilingual models that test on test data of per languages

For English, the multilingual model achieved a Macro F1 of 0.69961, which was slightly lower than its monolingual mDeBERTaV3-base counterpart. In contrast, for Italian, the multilingual model achieved a Macro F1 of 0.78619, which higher than both its monolingual mDeBERTaV3-base (0.77075) and UmBERTO (0.73341) performances. This indicates that multilingual training significantly benefits subjectivity detection in Italian.

For Arabic, the multilingual model achieved a Macro F1 of 0.68439, representing a substantial improvement over its monolingual mDeBERTaV3-base (0.57380) and AraELECTRA (0.59194) results. This is a critical finding, demonstrating that multilingual training can significantly boost performance for languages that are otherwise challenging in monolingual settings.

Bulgarian saw a Macro F1 of 0.72036, slightly lower than its monolingual mDeBERTaV3-base performance. For German, the multilingual model achieved a Macro F1 of 0.78946, which is very

close to its best monolingual performance obtained with German BERT base (0.78347).

5.3 Zero-shot Performance Analysis

This section presents the performance of the model in a zero-shot setting, where it is evaluated on languages not seen during training. The results are summarized in Table 6.

Languages	Macro F1	Precision	Recall	SUBJ F1
Polish	0.64251	0.94915	0.34783	0.50909
Greek	0.77467	0.61702	0.63043	0.62366
Romanian	0.72798	0.51852	0.80769	0.63158
Ukrainian	0.64025	0.45055	0.52564	0.48521

Table 6: Results for Zero-shot setting

These results highlight the multilingual model's ability to generalize to unseen languages, with varying degrees of success. Greek showed particularly strong performance than the other languages.

5.4 Overall performance Analysis

This section compares the performance of our model with the baseline and the best team results achieved in the 2025 Check That! Lab Task 1. The detailed comparison across monolingual, multilingual, and zero-shot scenarios is presented in Table 7. This comprehensive comparison indicates that our model consistently achieves better results for subjectivity detection across monolingual, multilingual, and zero-shot scenarios compared to the baseline. In comparison with the best-performing teams, our model demonstrates competitive performance, closely approaching top scores particularly for German, Italian, and multilingual contexts.

Languages	Our Result	Baseline	Best Team Result
English	0.71735	0.5370	0.8052
Italian	0.77075	0.6941	0.8104
Arabic	0.59194	0.5133	0.6884
Germany	0.78347	0.6960	0.8520
Multilingual	0.72472	0.6390	0.7550
Greek	0.77467	0.4159	0.5067
Polish	0.64251	0.5719	0.6922
Romanian	0.72798	0.6461	0.8126
Ukrainian	0.64025	0.6296	0.6424

Table 7: Comparing our best results in all three settings with baseline and best team result in this task.

6 Conclusion

Comparing our model's performance against the provided baseline, from the Lab page, F1 scores reveals a consistent advantage across all evaluated settings. In the monolingual setting, our model achieved significantly higher F1 scores for Italian (0.77075 vs. baseline 0.6941), Arabic (0.59194 with AraELECTRA vs. baseline 0.5133), German (0.78347 with German BERT base vs. baseline 0.6960), and English (0.71735 vs. baseline 0.5370). For the multilingual setting, our balanced approach yielded a Macro F1 of 0.72472, substantially outperforming the baseline's 0.6390. In the zero-shot setting, our model also demonstrated superior performance across all languages: Polish (0.64251 vs. baseline 0.5719), Ukrainian (0.64025 vs. baseline 0.6296), Romanian (0.72798 vs. baseline 0.6461), and Greek (0.77467 vs. baseline 0.4159). This comprehensive comparison indicates that our model consistently achieves better results for subjectivity detection across monolingual, multilingual, and zero-shot scenarios compared to the baseline. In the case of comparing our results with the other teams, we achieved performance placing us among the top-5 teams for most of the languages, except for English and Romanian. Despite these advancements, challenges remain, particularly with Arabic, which consistently showed lower performance, suggesting deeper linguistic complexities or data limitations. Future work should focus on advanced balancing techniques, detailed error analysis to understand specific linguistic hurdles, and exploring alternative model architectures or ensemble methods to further enhance performance and robustness in this task.

7 Links to external resources

Our codes are available at our GitHub repository.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Task 1: Subjectivity in News Articles. 2025. clef2025-checkthat-lab, task1.

Folkert Atze Leistra and Tommaso Caselli. 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.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

MDZ Digital Library team (dbmdz). 2025. dbmdz/bert-base-german-cased.