Subjectivity in News Articles

Professor: Torroni

Mehregan Nazarmohsenifakori

Shafagh Rastegari

Mohammad Pourtaheri

Problem Statement

Subjectivity detection helps identify bias and opinion in NLP tasks like fake news detection and sentiment analysis.

This project is address the problem in CLEF2025 challenge.

Our goal: Classify sentences as SUBJ or OBJ across 9 languages (some zero-shot).

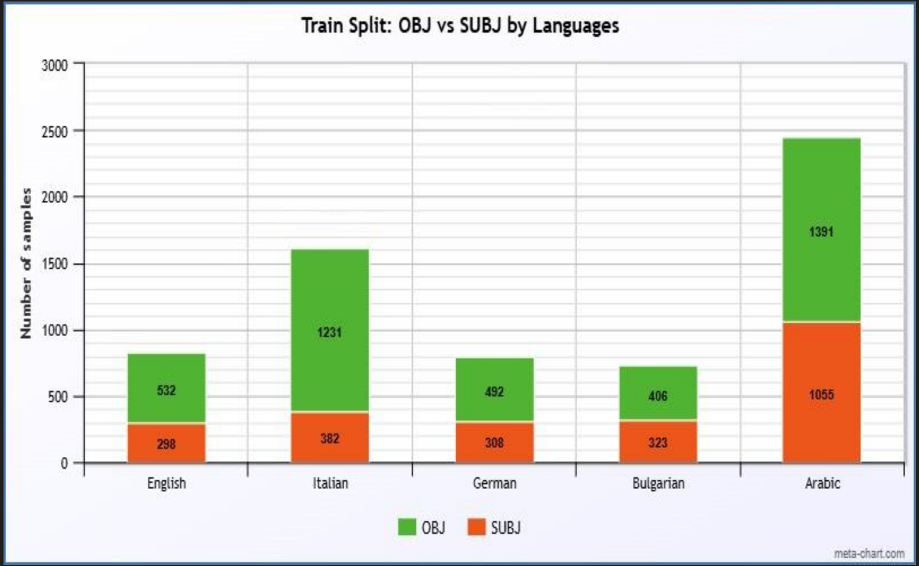
Challenges of the problem: Language diversity and Data imbalance.

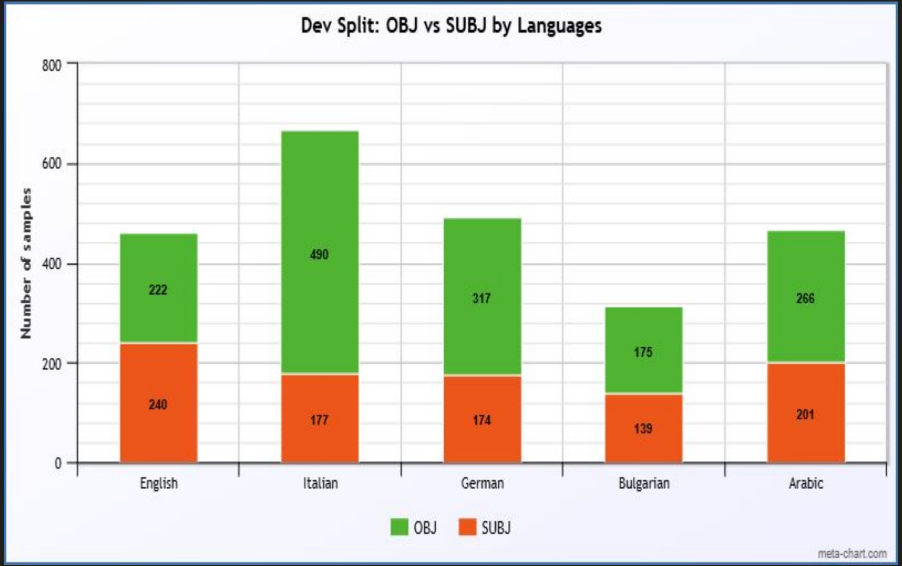
Dataset

There are 9 Languages:

5 with training data (English, Italian, German, Bulgarian, Arabic)

4 for zero-shot (Polish, Greek, Romanian, Ukrainian)





Baseline Model

Model: Logistic Regression on multilingual SBERT embeddings

Implementation: baseline.py script for training & prediction

Evaluation Metrics

Primary: Macro-averaged F1 (OBJ vs. SUBJ)

Secondary: Precision & Recall and F1 for the SUBJ class

Our Models

mDeBERTaV3-base

- Multilingual transformer used as the backbone
- Fine-tuned separately for each language (monolingual)
- Used also in multilingual and zero-shot settings

Language-Specific Models

German: BERT (dbmdz/bert-base-german-cased)

Italian: UmBERTo (Roberta-based Italian model)

Arabic: AraELECTRA (Electra-based Arabic model)

Experimental Settings

Monolingual

Train and test on data in a given language

Evaluate mDeBERTaV3 and language-specific models

Multilingual

Fine-tune **mDeBERTaV3** on 5-language:

Option 1: Combined dataset (all examples)

Option 2: Balanced dataset (60% OBJ / 40% SUBJ)

Zero-shot

Use best multilingual model (from balanced training)

Evaluate on unseen languages

No training data from these languages provided

Monolingual

Base Model

Fine-tuned mDeBERTaV3-base on: English, German, Italian, Arabic, Bulgarian
Pretrained on many languages, train and tested per language.

Language-Specific Models

German: German BERT (dbmdz)
Italian: UmBERTo
Arabic: AraELECTRA

Monolingual Result

➔ German BERT got better result than mDeBERTaV3 in German language

➔ mDeBERTaV3 outperformed UmBERTo in Italian

➔ Arabic had the lowest scores across models(Indicates Arabic is more challenging due to linguistic complexity or data imbalance)

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

Multilingual

Used mDeBERTaV3-base, trained on merged datasets from: English, German, Italian, Arabic, Bulgarian

Two training strategies:
- Combined Dataset: All data merged (imbalanced)
- Balanced Dataset: 60% OBJ / 40% SUBJ per language

Balanced training show better results than raw combined data by improving Precision and Macro F1 → importance of class balance in training multilingual models

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

Multilingual: Test Performance per Language

Now evaluate the best Multilingual model on each languages.

-English and Bulgarian: Monolingual model slightly better

-Italian and German: Multilingual model performed better

-Arabic: Significant improvement in Multilingual setting

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

Zero-Shot

Greek achieved the best zero-shot performance

Romanian had high recall, indicating strong sensitivity to subjective content.

Polish showed high precision, but very low recall.

Ukrainian had the lowest overall performance.

Performance varies widely, depends on linguistic similarity, model pretraining, and sentence structure overlap.

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

Hyperparameters

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

Our Model vs. Baseline vs. Best Team

All languages achieved better performance than the baseline

Our system ranked in the top 5 teams for most languages, including a 1st place in Greek and 2nd place in Arabic.

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

Conclusion

What we Achieved?

- Outperformed the baseline in all languages
- Ranked in the top 5 teams for most languages → 🥇 1st in Greek, 🥈 2nd in Arabic
- Strong multilingual generalization, including zero-shot settings
- Balanced training proved effective in improving overall precision and Macro F1

Remaining Challenges

- English and Romanian fell outside top 5 → Likely due to domain mismatch or pretraining limitations
- Arabic still remains the most challenging despite high ranking → Needs more robust models and error analysis

Future work

- Explore advanced balancing or augmentation techniques
- Try ensemble methods for more robust predictions

References

[1] Antoun, W., Baly, F., & Hajj, H. (2021). AraELECTRA: Pre-training text discriminators for Arabic language understanding. *WANLP 2021*.

[2] He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pretraining. *arXiv:2111.09543*.

[3] Leistra, F. A., & Caselli, T. (2023). Language-specific fine-tuning of mDeBERTaV3 for subjectivity detection. *CLEF Working Notes*.

[4] Parisi, L., Francia, S., & Magnani, P. (2020). UmBERTo: An Italian language model trained with whole word masking. *GitHub: musixmatchresearch/umberto*.

[5] DBMDZ team. (2025). bert-base-german-cased. *HuggingFace Model Repository*.

[6] CheckThat! Lab Task 1. (2025). Subjectivity in News Articles. *CLEF 2025 Evaluation Forum*.

[7] https://github.com/mehreganmohseni/Subjectivity-in-News-Articles

Thank you!

Contact our team

mehrega.nazarmohsenifakori@studio.unibo.it

shafagh.rastegari@studio.unibo.it

mohammad.pourtaheri@studio.unibo.it