SmolLab_SEU at CheckThat! 2025: How Well Do Multilingual Transformers Transfer Across News Domains for Cross-lingual Subjectivity Detection?

Notebook for the CheckThat! Lab at CLEF 2025

Md. Abdur Rahman[1,*], Md Al Amin[2], Md Sabbir Dewan[3,†], Md Jahid Hasan[4,†] and Md Ashiqur Rahman[1]

[1] *Southeast University, Bangladesh*
[2] *St. Francis College, Brooklyn, New York, USA*
[3] *Murdoch University, Perth, Australia*
[4] *Eastern University, Philadelphia, Pennsylvania, USA*

Abstract

Automated detection of subjectivity in news articles is an important problem for fighting against fake news and promoting journalistic accountability, but this is a challenging task in various linguistic settings. This paper shows our method on Task 1: Subjectivity of the CLEF 2025 CheckThat! Lab to identify objective information vs subjective opinion in news content available in several languages. To this aim, we experimented with a variety of Transformer models using architecture specifically designed for a particular language (GermanELECTRA-large, MARBERT-v2, RoBERTa-large) as well as multilingual (XLM-RoBERTa-large, mDeBERTaV3-base, InfoXLM-large) and zero-shot (mBERT-base) models. Our approach utilized the fine-tuning of pre-trained models with hyperparameters and class-weighted loss functions so as to tackle the imbalanced data. Experimental results show that our models perform well: German-ELECTRA-large achieved 0.8520 F1 in German, XLM-RoBERTa-large got 0.8356 F1 in Italian and 0.8040 in Romanian zero-shot, RoBERTa-large is best, with 0.7948 F1 on English, and InfoXLM-large achieves 0.7114 F1 on multilingual setting. In official ranking, our systems obtained 1st rank in Monolingual German, 2nd in Zero-shot Romanian, 3rd in Monolingual Italian, 5th in Zero-shot Ukrainian, 6th in Multilingual, 8th in Zero-shot Polish and 9th in Monolingual Arabic and English. Error analysis shows that monolingual models excelled monolingually, and multilingual architectures achieve better cross-lingual generalization in the zero-shot settings. This work provides insights into the suitability of Transformer in multilingual subjectivity detection and demonstrates the difficulties in recognizing subtle subjective cues in different linguistic environments.

Keywords

Subjectivity Detection, Transformer Models, Cross-lingual Transfer, Zero-shot Learning, Multilingual NLP, News Articles Analysis, Fine-tuning

1. Introduction

The rapid spread of news and opinions in the modern information-saturated environment shapes individual beliefs and societal debate to an unparalleled extent. The ability to separate objective facts from subjective claims is more important than ever given the growing frequency of subjective language in news reporting, particularly in the online media [1], and its regular connection with misleading information and fake news [2]. As a language tool, subjectivity captures the speaker's or writer's attitude, posture, and feelings toward the topic under discussion, therefore imprinting a personal mark

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

[*] Corresponding author.

[†] Authors contributed equally to this work.

2021200000025@seu.edu.bd (Md. A. Rahman); alaminhossine@gmail.com (M. A. Amin); sabbirdewan11@gmail.com (M. S. Dewan); mdjahid.hasan@eastern.edu (M. J. Hasan); ashiqur.rahman@seu.edu.bd (M. A. Rahman)
0009-0002-8595-5772 (Md. A. Rahman); 0009-0008-8981-9667 (M. A. Amin); 0009-0000-9332-0780 (M. S. Dewan); 0009-0001-1902-6142 (M. J. Hasan); 0009-0000-1391-3093 (M. A. Rahman)

on the communication [3]. Natural language processing (NLP) offers a powerful means to automate this difference, thus enhancing media literacy and supporting the campaign against misinformation.

This paper details our participation in Task 1: Subjectivity of the CLEF 2025 CheckThat! Lab [4, 5, 6]. The work tests algorithms for determining whether news article sentences fall into objective (OBJ) or subjective (SUBJ) classes. Using datasets [7, 8] in Arabic, Bulgarian, English, German, and Italian for training and development, it spans monolingual, multilingual, and zero-shot assessment scenarios with additional undiscovered languages (Greek, Polish, Romanian, Ukrainian) incorporated in the test phase. The macro-averaged F1-score is the formal benchmark for performance. Section 3 gives more details on the work and the datasets.

Our basic approach involves fine-tuning a wide range of pre-trained Transformer-based language models [9], specifically chosen and modified for each language or context. We examined strong multilingual designs as well as language-specific models.

Our main contributions consist of the following:

- We evaluated a varied suite of Transformer-based models for subjectivity detection across monolingual, multilingual, and zero-shot settings.
- We investigated model performance holistically on the subjectivity task, providing information on their capacities, cross-lingual transfer efficacy, and common error patterns over several languages and settings.

2. Related Works

The problem of detecting subjective expressions in text, commonly known as subjectivity detection, is a well-established research area in Natural Language Processing (NLP) and is closely related to sentiment analysis. Wiebe et al. [10] were among the early researchers to annotate subjectivity in text data. They have made their own tool called OpinionFinder [11], which tried to use different lexical and syntactic indicators to identify and analyze opinions. The task has recently been addressed successfully by using various deep-learning methods, leading to impressive performances. After preprocessing and annotating a political and ideological discussion dataset using a unique mix of lexicon-based and syntactic pattern-based approaches, Al Hamoud et al. [12] conducted a sentence subjectivity analysis on it. GloVe word embeddings enable them to evaluate six deep learning models (LSTM, BiLSTM, GRU, BiGRU, LSTM with attention, and BiLSTM with attention). On their two-class subjectivity classification problem, their best-performing model, LSTM with attention, achieved an accuracy of 97.39%, an F1-score of 99.20%, and a Kappa coefficient of 94.76%.

For joint polarity and subjectivity detection, Satapathy et al. [13] presented a knowledge-sharing-based multitask learning (MTL) paradigm. They linked the two tasks using a neural tensor network (NTN) and BERT embeddings. Their MTL framework, which combines BERT and NTN (MTLshared-NTN), achieved an accuracy of 95.1% for subjectivity detection and 94.6% for polarity detection on movie review datasets.

ThatiAR, a sizable manually annotated Arabic news sentence dataset for subjectivity detection, was first presented by Suwaileh et al. [8], along with GPT-4-generated explanations and directions for LLM fine-tuning. They benchmarked several large language models (LLMs) and Pre-trained Language Models (PLMs). On their ThatiAR data, GPT-4 with 3-shot in-context learning attained an F1-score of 0.800 for the “Subjective” class.

For the CLEF 2023 CheckThat! Lab, Pachov et al. [14] developed a method for subjectivity identification in English news stories. Combining three separate techniques, a fine-tuned sentence embedding encoder with dimensionality reduction (SBERT, PCA, ElasticNet), a few-shot learning SetFit model, and a fine-tuned `xlm-roberta-base` model and their best-performing strategy was a simple majority voting ensemble. The last ensemble on the English exam set achieved an overall macro F1-score of 0.77, scoring 0.77 for the “subjective” class and 0.78 for the “objective” class.

Using Google’s pre-trained large language model (LLM), Gemini, Gruman and Kosseim [15] detailed their zero-shot method for subjectivity categorization (Task 2) at CLEF 2024. Their approach consisted of fast engineering, in which a random portion of training data in the input prompt for context and augmenting the test data by producing two paraphrases; a majority vote over the original and paraphrased words selected the final label. Their F1 score on the English test set for Task 2-Subjectivity was 0.370.

3. Task and Dataset Description

We took part in Task 1: Subjectivity of the CLEF 2025 CheckThat! Lab [4, 5, 6]. The main goal of the task is to distinguish sentences taken from news articles between subjective (SUBJ), reflecting the opinion of the author, and objective (OBJ), provisioning a factual information. This binary classification problem is organized into three distinct settings: monolingual (training/testing with one language), multilingual (training/testing with a combined dataset of languages), and zero-shot (training with different languages and testing with unseen languages).

The official training dataset [7, 8] is a set of sentences from news articles. Training and development set splits are available for five languages: Arabic, Bulgarian, English, German, and Italian. Table 1 summarizes the number of sentences with SUBJ / OBJ counts per split for the five languages. We also have test sets in Greek, Polish, Romanian and Ukrainian (mainly zero-shot) and one multilingual test set. For multilingual and zero-shot setting, data can be pooled across all language-specific languages. The overall system performance is measured by the macro-average F1-score over the SUBJ and OBJ classes.

Table 1
Number of sentences per split and language, where SUBJ stands for subjective and OBJ stands for objective.

Language	Split	Sentences	OBJ	SUBJ
Arabic	Train	2446	1391	1055
	Dev	467	266	201
	Test	1036	727	309
English	Train	830	532	298
	Dev	462	222	240
	Test	300	215	85
German	Train	800	492	308
	Dev	491	317	174
	Test	347	229	118
Italian	Train	1613	1231	382
	Dev	667	490	177
	Test	299	192	107
Bulgarian	Train	691	379	312
	Dev	306	167	139
Multilingual	Test	1982	1363	619
Greek	Test	284	236	46
Polish	Test	351	190	161
Romanian	Test	206	154	52
Ukrainian	Test	297	219	78

4. Methodology

This section describes the methods used for Task 1: Subjectivity, which includes monolingual, multilingual, and zero-shot setting. Our main methodology consists of fine-tuning a diverse set of pre-trained Transformer-based language models, adapted to either a language or setting, as illustrated in Figure 1. The main objective was efficient classification of sentences into subjective (SUBJ) and objective (OBJ) while maximizing the macro F1-score.

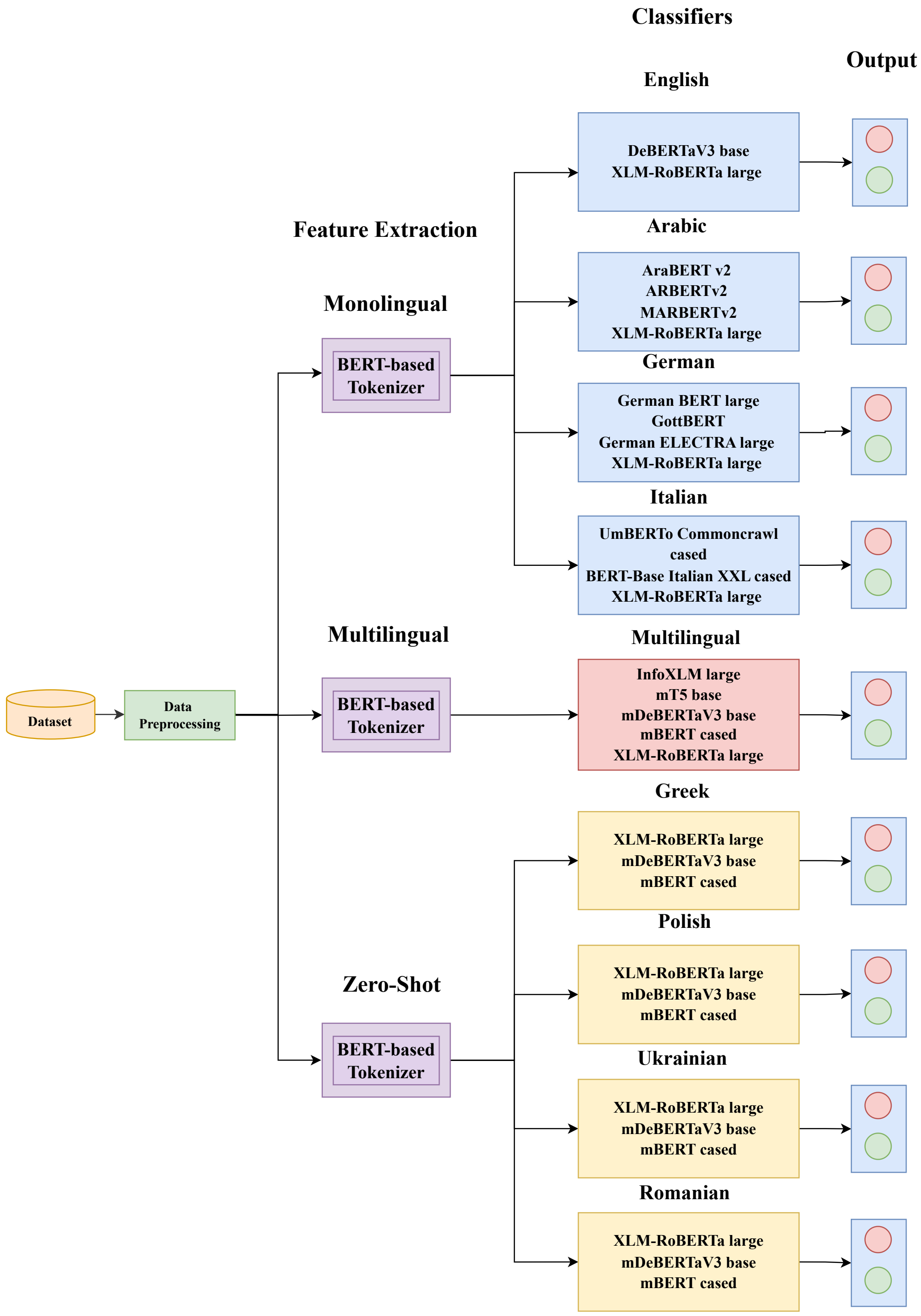


Figure 1: Overview of our methodology for Task 1: Subjectivity classification.

4.1. Data Preprocessing and Feature Extraction

We created a data preprocessing pipeline that first loads the TSV datasets available for each language and data split (train, development, and test). A standard text cleaning procedure was adopted, which largely involved the removal of leading and trailing whitespaces. We also discarded any rows from the dataset that were missing either the sentence text or its corresponding label. For feature extraction, we used pre-trained, model-specific tokenizers using `AutoTokenizer` from the Hugging Face Transformers library[1] for the range of different pre-trained architectures across our experiments (i.e., RoBERTa, DeBERTa, XLM-RoBERTa, language-specific BERTs, and mT5). Sentences were tokenized into subword units. This returned the `input_ids` and `attention_mask` tensors, which are the direct input features for the Transformer models. These sequences were tokenized by padding shorter sequences and truncating longer sequences to a max length of 128 tokens. Special tokens (`[CLS]`, `[SEP]`) specific to each Transformer model were automatically added by the tokenizer. Finally, the vocabulary `'SUBJ'` and `'OBJ'` were indexed sequentially as 1 and 0 for training and evaluation.

4.2. Transformer-Based Models

The core of our experimental methodology for all types of experimental settings, monolingual, multilingual and zero-shot, was fine-tuning of pre-trained Transformer models [9]. These models, in particular famous for their self-attention mechanisms, have a great potential to capture intricate contextual nuances and long-range dependencies in text, making them extremely powerful for a variety of Natural Language Processing (NLP) tasks [16, 17]. An overview of the general experimental pipeline, demonstrating the process from raw data ingestion and pre-processing, tokenization, feature extraction and subsequent classification by a variety of Transformer model species designed for the task setting, is given in Fig. 1. For each Transformer model, a classifier head was attached to the pre-trained encoder stack. This head generally consisted of a dropout layer for regularization and then a final linear layer, which projected the context-sensitive representation of the [CLS] token's final hidden state (or equivalent pooled output representation based on the architecture of the model) into a logit space of the two target classes: subjective (SUBJ) and objective (OBJ).

In the monolingual experiments, we concentrated on tuning models on individual language-specific corpora. In the case of English, we performed our experiments using RoBERTa-large [18], which is more robust to pre-training, and DeBERTaV3-base [19], which is equipped with advancements such as disentangled attention. The analysis of subjectivity in Arabic involved utilizing Arabic-pretrained models such as AraBERT v2 [20], ARBERTv2 and MARBERTv2 [21] all specifically pre-trained on large Arabic corpora, in addition to the widely multilingual model XLM-RoBERTa large [22]. For German, we tried German BERT large [23], GottBERT [24] (trained on historical and modern German), German ELECTRA large [23] (with replaced token detection as a pre-training objective) as well as XLM-RoBERTa large. Our experiments on Italian were performed with UmBERTo Commoncrawl cased[2] and BERT-Base Italian XXL cased[3], Italian-tuned transformers, and XLM-RoBERTa large. It was a choice that sought to benefit from not only language-specific optimization but also the strong multilingual representations in monolingual training.

In the multilingual setting, our approach consisted of training models on an aggregated dataset created by concatenating the training data from all five main languages: Arabic, Bulgarian, English, German, and Italian. This approach was proposed so the models could learn mutual, language-independent cues for subjectivity. For this task, we used pre-trained multilingual models that were chosen specifically for their multilingual capacity, which are XLM-RoBERTa large, InfoXLM large [25] (that includes cross-lingual pre-training objectives like Translation Language Modeling), mT5 base [26] (modified from its initial sequence-to-sequence architecture for this classification task), mDeBERTaV3 base, and mBERT cased [16].

The zero-shot scenario clearly measured the multilingual model's generalization capability. Models were first pre-trained on the joint multilingual dataset (Arabic, Bulgarian, English, German and Italian). We then test the tuned models using test sets of unseen test languages (Greek, Polish, Ukrainian and Romanian) without any fine-tuning in between. This configuration profoundly questions the extent to which models are able to generalize the learned subjectivity properties across different linguistic contexts based exclusively on the cross-lingual information that they have gathered during pre-training

[1] https://huggingface.co/transformers/

[2] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[3] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

Table 2
Key hyperparameters for Transformer-based models. LR: Learning Rate, WD: Weight Decay, BS: Batch Size, EP: Epochs (Initial/Final Training), Max. Len: Maximum Sequence Length, WU Prop.: Warmup Proportion. For Zero-Shot, models were trained on the “All Seen” multilingual data with the listed hyperparameters and then evaluated on the target unseen language.

Target/Setting	Model	LR	WD	BS	EP	Max. Len	WU Prop.
Monolingual: English							
English	roberta-large	2.0e-5	0.01	16	5/3	128	0.1
English	DeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Monolingual: Arabic							
Arabic	AraBERT-v2	1.8e-5	0.01	16	5/3	128	0.1
Arabic	ARBERT-v2	1.8e-5	0.01	16	5/3	128	0.1
Arabic	MARBERT-v2	1.8e-5	0.01	16	5/3	128	0.1
Arabic	XLM-RoBERTa-large	2.0e-5	0.01	16	5/3	128	0.1
Monolingual: German							
German	German-BERT-large	1.5e-5	0.01	16	5/3	128	0.1
German	GottBERT	1.5e-5	0.01	16	5/3	128	0.1
German	German-ELECTRA-large	1.5e-5	0.01	16	5/3	128	0.1
German	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Monolingual: Italian							
Italian	UmBERTo-Commoncrawl-cased	1.5e-5	0.01	16	5/3	128	0.1
Italian	BERT-Base-Italian-XXL-cased	1.8e-5	0.01	16	5/3	128	0.1
Italian	XLM-RoBERTa-large	1.8e-5	0.01	16	5/3	128	0.1
Multilingual (All Seen → Test Set)							
Multilingual	InfoXLM-large	1.5e-5	0.01	16	5/3	128	0.1
Multilingual	mT5-base	1.4e-5	0.01	8	5/3	128	0.1
Multilingual	mDeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Multilingual	mBERT-cased	1.8e-5	0.01	16	5/3	128	0.1
Multilingual	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Zero-Shot: Greek							
Greek	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Greek	mDeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Greek	mBERT-cased	1.8e-5	0.01	16	5/3	128	0.1
Zero-Shot: Polish							
Polish	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Polish	mDeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Polish	mBERT-cased	1.8e-5	0.01	16	5/3	128	0.1
Zero-Shot: Ukrainian							
Ukrainian	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Ukrainian	mDeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Ukrainian	mBERT-cased	1.8e-5	0.01	16	5/3	128	0.1
Zero-Shot: Romanian							
Romanian	XLM-RoBERTa-large	1.5e-5	0.01	16	5/3	128	0.1
Romanian	mDeBERTaV3-base	1.8e-5	0.01	16	5/3	128	0.1
Romanian	mBERT-cased	1.8e-5	0.01	16	5/3	128	0.1

and multilingual fine-tuning. We mostly investigated the efficiency of zero-shot transfer using XLM-RoBERTa large, mDeBERTaV3 base and mBERT cased architectures.

The fine-tuning procedure was mostly the same throughout all experiments. We used the AdamW optimizer [27] that provides an enhanced weight decay regularization compared to the original Adam. A linear learning rate warm-up and decay scheduler was adopted to stabilize the training process, with the warm-up phase set as the first 10% steps regarding the total training steps. The loss function was the standard Cross Entropy Loss, as natively implemented within the Hugging Face AutoModelForSequenceClassification classes. For some models, if the class imbalance was high in the training set of a language (e.g., RoBERTa-large for English), pre-calculated class weights were supplied to the model, enabling the framework to internally adjust the loss computation to give more importance to underrepresented classes. During training, models were first trained on the specified training sets, and their performance was measured on the corresponding development sets using the macro F1-score. This optimal checkpoint was then subjected to a subsequent, shorter phase of fine-tuning (3 epochs) on a combined dataset comprising both the original training and development data, often employing a reduced learning rate. This last recalibrated model was then applied to predict the specific test sets. All experiments were performed with a set of fixed random seeds (42) for reproducibility. A max-norm of 1.0 was used to clip gradients to avoid exploding during backpropagation. Hyperparameter settings for the main models used in each experiment are detailed in Table 2. All experiments were conducted in a Kaggle environment with an NVIDIA Tesla P100 GPU, which provided the necessary computational resources to fine-tune large-scale models.

5. Result Analysis

This section provides a comprehensive experimental analysis of our model's performance for the task of subjectivity classification in typical monolingual, multilingual and zero-shot settings. The macro-averaged F1 is the official evaluation metric. Table 3 presents the recall, precision and F1-score of Transformer-based models on the corresponding test set.

In the monolingual setting, for English, RoBERTa-large (0.7948 F1) outperformed DeBERTaV3-base. In Arabic, MARBERT-v2 (0.5885 F1) achieved the best performance, closely followed by XLM-RoBERTa large (0.5747 F1), outperforming AraBERTv2 and ARBERTv2. The German ELECTRA large (0.8520 F1) yielded the best performance for German. Its success can be attributed to its pre-training on a massive German-only corpus combined with the efficient "replaced token detection" objective, allowing it to capture language-specific nuances of subjectivity more effectively than generalist multilingual models. Conversely, in Italian, the multilingual XLM-RoBERTa large (0.8356 F1) slightly outperformed the Italian-specific BERT-Base XXL Cased (0.8309 F1). This indicates that the sheer scale and diversity of XLM-RoBERTa's pre-training can sometimes create representations robust enough to compete with, or even exceed, those of large models trained on a single language's corpus.

In the multilingual setting, where models were trained on data from all five languages, InfoXLM-large (0.7114 F1) and mDeBERTaV3-base (0.7109 F1) performed best. The strong performance of InfoXLM is particularly noteworthy; its pre-training includes cross-lingual objectives like Translation Language Modeling (TLM), which explicitly forces the model to align sentence representations across languages. This is a powerful advantage for learning a language-independent concept like subjectivity from a mixed-corpus. XLM-RoBERTa large also achieved strong results (0.6993 F1), leveraging its vast multilingual vocabulary, while the older mBERT-cased and sequence-to-sequence mT5-base models were notably behind.

The zero-shot setting tested generalization to unseen languages. Here, the architectural strengths and pre-training data of the models become paramount. XLM-RoBERTa large excelled on Ukrainian (0.6730 F1) and Romanian (0.8040 F1). This success is likely due to its massive pre-training on 100 languages from CommonCrawl, which includes both Slavic (related to Ukrainian) and Romance (related to Romanian) languages. This linguistic proximity allows the model to effectively transfer learned subjectivity patterns to unseen but related languages. In contrast, mDeBERTaV3-base proved superior for Greek (0.4945 F1) and Polish (0.5737 F1). Its architectural improvements, such as disentangled attention, may allow it to capture more abstract syntactic and semantic cues of subjectivity that generalize better across more distant language families. These results underscore that successful zero-shot transfer depends not only on multilingual exposure but also on the specific relationship between the source and target languages.

While specialized monolingual models can achieve peak performance in their native language, our

Table 3
Model performance (Recall, Precision, F1-score) across Monolingual, Multilingual, and Zero-shot settings on test data.

Setting	**Language**	**Model**	**Recall**	**Precision**	**F1**
		Monolingual			
Monolingual	**English**	**RoBERTa-large**	**0.7948**	**0.7948**	**0.7948**
		DeBERTaV3-base	0.7299	0.7362	0.7328
	Arabic	**MARBERT-v2**	**0.5871**	**0.5905**	**0.5885**
		AraBERT-v2	0.5070	0.5077	0.5052
		ARBERT-v2	0.5494	0.5567	0.5462
		XLM-RoBERTa-large	0.5807	0.5725	0.5747
	German	**German-ELECTRA-large**	**0.8622**	**0.8442**	**0.8520**
		German-BERT-large	0.8318	0.7996	0.8117
		GottBERT	0.7818	0.7463	0.7582
		XLM-RoBERTa-large	0.8282	0.7639	0.7822
	Italian	**XLM-RoBERTa-large**	**0.8270**	**0.8461**	**0.8356**
		BERT-Base-Italian-XXL-Cased	0.8183	0.8493	0.8309
		UmBERTo-Commoncrawl-Cased	0.7620	0.8049	0.7750
		Multilingual			
Multilingual	**All Seen**	**InfoXLM-large**	**0.7743**	**0.6935**	**0.7114**
		XLM-RoBERTa-large	0.7148	0.6906	0.6993
		mDeBERTaV3-base	0.7292	0.7010	0.7109
		mBERT-cased	0.6286	0.6314	0.6299
		mT5-base	0.5510	0.5059	0.4360
		Zero-shot			
Zero-shot	**Greek**	**mDeBERTaV3-base**	**0.5058**	**0.4860**	**0.4945**
		XLM-RoBERTa-large	0.5754	0.3826	0.3924
		mBERT-cased	0.4140	0.4376	0.4225
	Polish	**mDeBERTaV3-base**	**0.7803**	**0.6189**	**0.5737**
		XLM-RoBERTa-large	0.7791	0.5905	0.5268
		mBERT-cased	0.7772	0.6127	0.5643
	Ukrainian	**XLM-RoBERTa-large**	**0.6791**	**0.6682**	**0.6730**
		mDeBERTaV3-base	0.6202	0.6294	0.6237
		mBERT-cased	0.5958	0.6070	0.5990
	Romanian	**XLM-RoBERTa-large**	**0.8397**	**0.7817**	**0.8040**
		mDeBERTaV3-base	0.7802	**0.8004**	0.7891
		mBERT-cased	0.7492	0.7523	0.7507

findings consistently show that large, robustly pre-trained multilingual models like XLM-RoBERTa-large and mDeBERTaV3-base are top performers across all settings. Their strength lies in their ability to learn universal linguistic patterns. XLM-RoBERTa's advantage comes from its vast, 100-language pre-training corpus, making it a powerful baseline for cross-lingual transfer. Meanwhile, models like InfoXLM and mDeBERTaV3 leverage more advanced pre-training objectives (TLM) or architectural designs (disentangled attention) to further enhance this generalization capability.

5.1. Inference Time and Resource Constraints

To assess the practical viability of our approach, particularly in zero-shot settings with potential resource constraints, we measured the inference time of our top-performing zero-shot models. Inference was performed on a single NVIDIA Tesla P100 GPU. We measured the total time required to process the entire test set for each unseen language, and the results are presented in Table 4.

Table 4
Inference time analysis for the best-performing models in the zero-shot setting.

Language	Model	Test Examples	Total Time (s)	Avg. Time/Example (ms)
Greek	mDeBERTaV3-base	351	2.2	~6.3
Polish	mDeBERTaV3-base	284	1.9	~6.7
Ukrainian	XLM-RoBERTa-large	297	4.7	~15.8
Romanian	XLM-RoBERTa-large	206	2.7	~13.1

The analysis highlights a clear trade-off between model performance and computational cost. The larger model, `XLM-RoBERTa-large`, while achieving superior F1-scores on Ukrainian and Romanian, is more than twice as slow as the `mDeBERTaV3-base` model. Specifically, `XLM-RoBERTa-large` required approximately 13-16 ms per sentence, whereas the base-sized `mDeBERTaV3-base` was significantly faster at around 6-7 ms per sentence. This disparity is a direct result of model size and complexity; larger models demand more GPU memory and compute cycles, leading to higher latency. This is a critical factor for deployment in production systems where low latency is often a firm requirement. For scenarios with tight resource constraints, a smaller model like `mDeBERTaV3-base` could be a more pragmatic choice, despite a potential drop in performance for certain languages.

6. Error Analysis

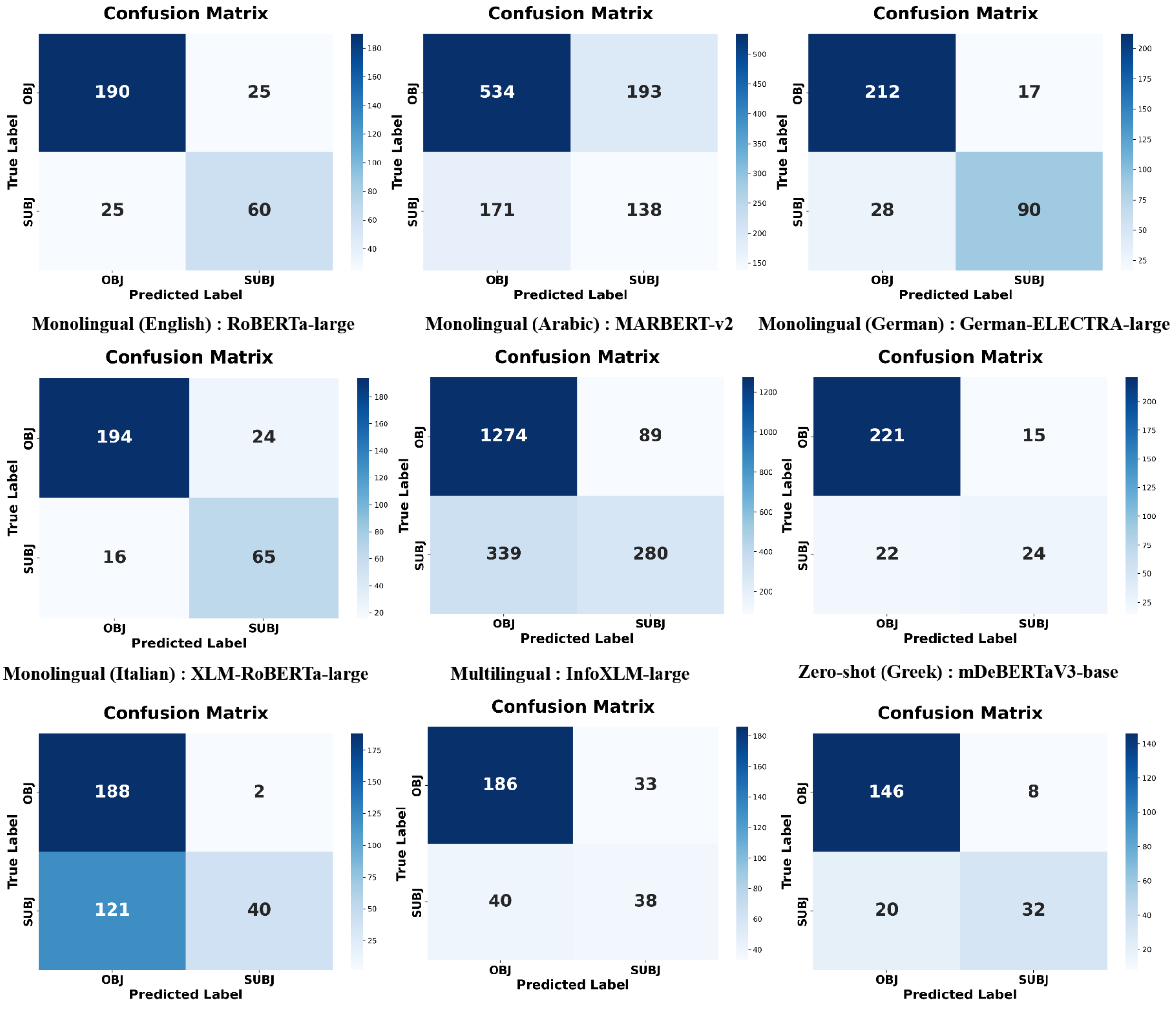
We conducted both quantitative and qualitative error analyses of the models to gain a deeper understanding of the reasoning behind model behaviour and identify potential avenues for improvement in each model.

6.1. Quantitative Analysis

Figure 2 shows an aggregation of confusion matrices for our best models on individual monolingual sets, the best multilingual model, and the best zero-shot model for each of the previously unseen target languages. In monolingual settings, models such as German-ELECTRA-large (German) or RoBERTa-large (English) demonstrated good performance with less biased confusion between Objective (OBJ) and Subjective (SUBJ) classes. A very modest misclassification bias of SUBJ as OBJ was found in English RoBERTa-large (25 False Negatives for SUBJ vs 25 False Positives for OBJ). For Arabic, MARBERT-v2 had more problems identifying the subjective sentences, with 171 SUBJ instances classified as OBJ versus 193 OBJ instances classified as SUBJ. In Italian, XLM-RoBERTa-large also showed a slight asymmetry, with more SUBJ instances misclassified as OBJ.

Our top-performing multilingual model, InfoXLM-large, exhibited a clear tendency to make more frequent mistakes in mixing cases than in separating them (339 False Negatives as SUBJ and 89 as OBJ), which suggested that the infused information in this model leaned towards separating SUBJ from OBJ. This implies that if trained on various linguistic inputs, the model could take a more conservative position, where it defaults to an objective reading for ambiguous sentences.

Error rates were generally higher for the zero-shot cases. For Greek, mDeBERTaV3-base made more classification errors involving instances of SUBJ misclassified as OBJ. There was a strong bias to predict OBJ in Polish under mDeBERTaV3-base (top-left, bottom row) in which 121 of SUBJ sentences were categorized as OBJ, while there were only 2 errors due to OBJ as SUBJ. Ukrainian and Romanian,



Monolingual (English) : RoBERTa-large

Monolingual (Arabic) : MARBERT-v2

Monolingual (German) : German-ELECTRA-large

Monolingual (Italian) : XLM-RoBERTa-large

Multilingual : InfoXLM-large

Zero-shot (Greek) : mDeBERTaV3-base

Zero-shot (Polish) :mDeBERTaV3-base

Zero-shot (Ukrainian) : XLM-RoBERTa-large

Zero-shot (Romanian) : XLM-RoBERTa-large

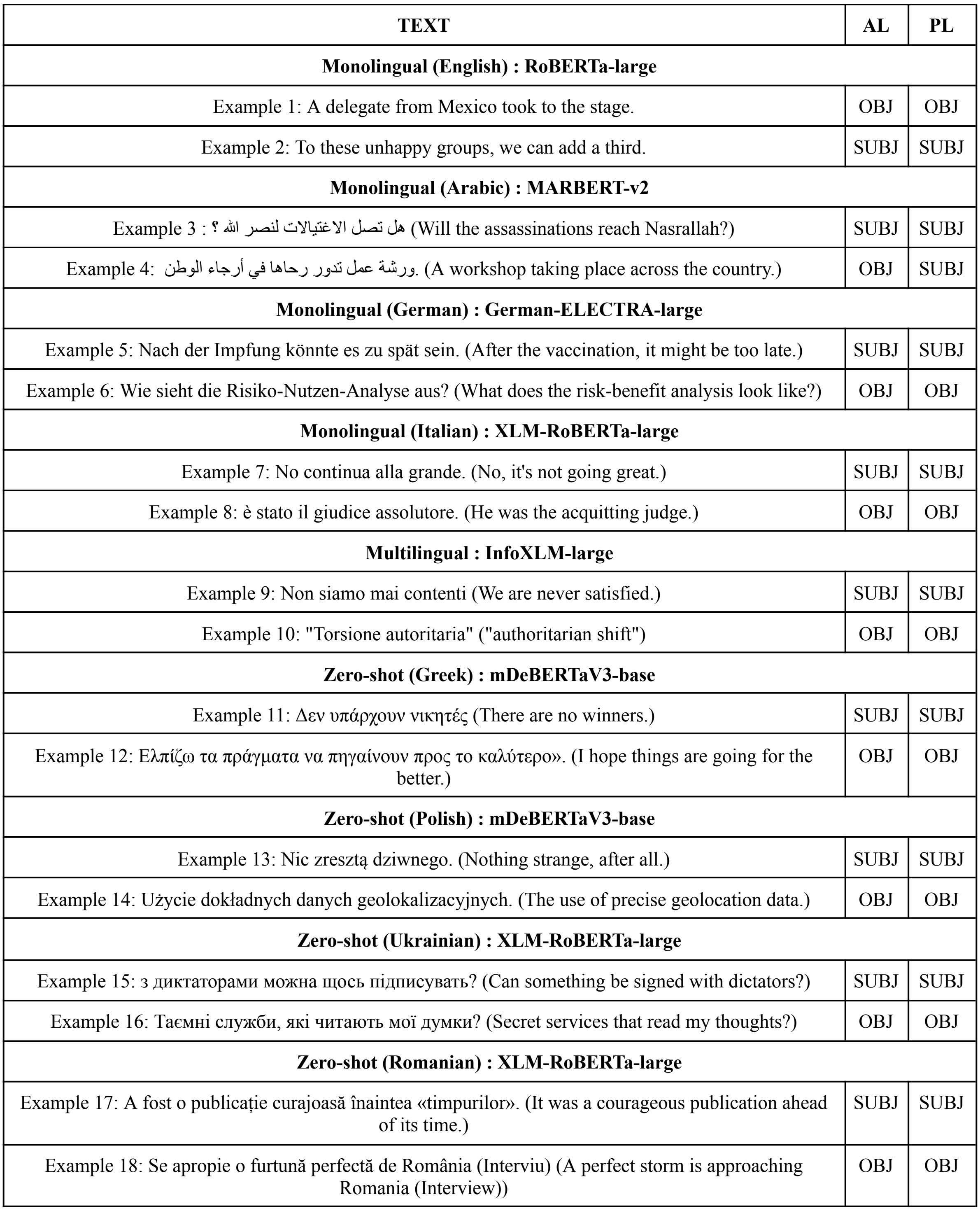
Figure 2: Composite of confusion matrices for the best performing models in each setting/language. Rows (top to bottom) generally represent: Monolingual (English, Arabic, German, Italian) & Multilingual & Zero-shot (Greek, Polish, Ukrainian, Romanian). OBJ: Objective, SUBJ: Subjective.

evaluated with XLM-RoBERTa-large, also showed a tendency to misclassify subjective statements as objective, albeit less pronounced than in Polish.

6.2. Qualitative Analysis

Figure 3 shows prediction examples and their corresponding actual labels. Here are several simple statements of fact (example 1 - English) and unambiguously opinionated sentences (example 5 - German) that were accurately discriminated in both groups of settings.

Mistakes often occur with mild or subtle notes of subjectivity. Consider the zero-shot Greek example in instance 12, which conveys a subjective hope but was classified as OBJ by mDeBERTaV3-base. While the true label is also OBJ, this case may reflect annotation noise or highlight the model's difficulty in handling such nuanced expressions. It is clear that the model has a strong bias toward the OBJ in zero-shot Polish and correctly identifies example 14. Short and possibly biased phrases, such as example 10, were correctly recognized as OBJ by InfoXLM-large in the multilingual condition, likely due to contextual training. Final sentences that use rhetorical questions or permeate with hidden subjectivity, as the Ukrainian example 16, which are annotated as OBJ but may have some subjective weight, demonstrate that even difficult tasks are an issue for state-of-the-art models. These findings also suggest that future research may want to focus on enhancing models' ability to utilize subtle linguistic



TEXT	AL	PL
Monolingual (English) : RoBERTa-large		
Example 1: A delegate from Mexico took to the stage.	OBJ	OBJ
Example 2: To these unhappy groups, we can add a third.	SUBJ	SUBJ
Monolingual (Arabic) : MARBERT-v2		
Example 3 : هل تصل الاغتيالات لنصر الله ؟ (Will the assassinations reach Nasrallah?)	SUBJ	SUBJ
Example 4: ورشة عمل تدور رحاها في أرجاء الوطن. (A workshop taking place across the country.)	OBJ	SUBJ
Monolingual (German) : German-ELECTRA-large		
Example 5: Nach der Impfung könnte es zu spät sein. (After the vaccination, it might be too late.)	SUBJ	SUBJ
Example 6: Wie sieht die Risiko-Nutzen-Analyse aus? (What does the risk-benefit analysis look like?)	OBJ	OBJ
Monolingual (Italian) : XLM-RoBERTa-large		
Example 7: No continua alla grande. (No, it's not going great.)	SUBJ	SUBJ
Example 8: è stato il giudice assolutore. (He was the acquitting judge.)	OBJ	OBJ
Multilingual : InfoXLM-large		
Example 9: Non siamo mai contenti (We are never satisfied.)	SUBJ	SUBJ
Example 10: "Torsione autoritaria" ("authoritarian shift")	OBJ	OBJ
Zero-shot (Greek) : mDeBERTaV3-base		
Example 11: Δεν υπάρχουν νικητές (There are no winners.)	SUBJ	SUBJ
Example 12: Ελπίζω τα πράγματα να πηγαίνουν προς το καλύτερο». (I hope things are going for the better.)	OBJ	OBJ
Zero-shot (Polish) : mDeBERTaV3-base		
Example 13: Nic zresztą dziwnego. (Nothing strange, after all.)	SUBJ	SUBJ
Example 14: Użycie dokładnych danych geolokalizacyjnych. (The use of precise geolocation data.)	OBJ	OBJ
Zero-shot (Ukrainian) : XLM-RoBERTa-large		
Example 15: з диктаторами можна щось підписувать? (Can something be signed with dictators?)	SUBJ	SUBJ
Example 16: Таємні служби, які читають мої думки? (Secret services that read my thoughts?)	OBJ	OBJ
Zero-shot (Romanian) : XLM-RoBERTa-large		
Example 17: A fost o publicație curajoasă înaintea «timpurilor». (It was a courageous publication ahead of its time.)	SUBJ	SUBJ
Example 18: Se apropie o furtună perfectă de România (Interviu) (A perfect storm is approaching Romania (Interview))	OBJ	OBJ

Figure 3: Examples of model predictions (PL) versus actual labels (AL) for selected models and languages.

cues for subjectivity, particularly in cross-lingual and low-resource settings, and addressing model biases towards majority classes observed in some zero-shot transfers.

7. Conclusion

In this paper, we present our contribution to the CLEF 2025 CheckThat! Lab Task 1: Subjectivity. We conducted an extensive comparison of monolingual, multilingual, and zero-shot Transformer-based models. Our results reveal a clear trade-off: while language-specific models (e.g., German ELECTRA large) excel on monolingual tasks by capturing local linguistic nuances, robust multilingual models show superior versatility and transferability. The strong performance of XLM-RoBERTa-large and mDeBERTaV3-base in multilingual and zero-shot settings stems from their underlying design. We attribute XLM-RoBERTa's success to its massive 100-language pre-training data, which fosters generalization to related languages (e.g., Romanian, Ukrainian). Meanwhile, mDeBERTaV3's advanced architecture appears to help in transferring more abstract patterns to less related languages (e.g., Greek, Polish). Error analysis confirmed that challenges with subtle subjectivity and majority-class bias persist, especially in zero-shot scenarios. Our future work will focus on developing more effective fine-tuning techniques, addressing data imbalance, and improving cross-lingual generalization. Furthermore, we plan to explore model optimization techniques, such as knowledge distillation and quantization, to better balance the trade-off between predictive performance and the computational efficiency highlighted in our analysis.

Acknowledgments

This work was supported by Southeast University, Bangladesh.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly for grammar and spelling checks. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. S. Blake, et al., News in a digital age: Comparing the presentation of news information over time and across media platforms, Rand Corporation, 2019.

[2] C. L. M. Jeronimo, L. B. Marinho, C. E. Campelo, A. Veloso, A. S. da Costa Melo, Fake news classification based on subjective language, in: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, 2019, pp. 15–24.

[3] L. Feng, On the subjectivity and intersubjectivity of language, Communication and Linguistics Studies 6 (2020) 1–5.

[4] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[5] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[6] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, G. Razvan, A. Mihail, Overview of the CLEF-2025 CheckThat! lab task 1

on subjectivity in news article, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[7] F. Antici, F. Ruggeri, A. Galassi, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on English news articles, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 273–285. URL: https://aclanthology.org/2024.lrec-main.25/.

[8] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, Thatiar: subjectivity detection in arabic news sentences, arXiv preprint arXiv:2406.05559 (2024).

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] J. Wiebe, R. Bruce, T. P. 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.

[11] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: A system for subjectivity analysis, in: Proceedings of HLT/EMNLP 2005 interactive demonstrations, 2005, pp. 34–35.

[12] A. Al Hamoud, A. Hoenig, K. Roy, Sentence subjectivity analysis of a political and ideological debate dataset using lstm and bilstm with attention and gru models, Journal of King Saud University-Computer and Information Sciences 34 (2022) 7974–7987.

[13] R. Satapathy, S. R. Pardeshi, E. Cambria, Polarity and subjectivity detection with multitask learning and bert embedding, Future Internet 14 (2022) 191.

[14] G. Pachov, D. Dimitrov, I. Koychev, P. Nakov, Gpachov at checkthat! 2023: a diverse multi-approach ensemble for subjectivity detection in news articles, arXiv preprint arXiv:2309.06844 (2023).

[15] S. Gruman, L. Kosseim, Clac at checkthat! 2024: a zero-shot model for check-worthiness and subjectivity classification, Faggioli et al.[22] (2024).

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.

[18] 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).

[19] 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).

[20] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, H. Mubarak (Eds.), Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 9–15. URL: https://aclanthology.org/2020.osact-1.2/.

[21] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, ARBERT & MARBERT: Deep bidirectional transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7088–7105. URL: https://aclanthology.org/2021.acl-long.551. doi:10.18653/v1/2021.acl-long.551.

[22] 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 (2019).

[23] B. Chan, S. Schweter, T. Möller, German's next language model, arXiv preprint arXiv:2010.10906 (2020).

[24] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, M. Boeker, Gottbert: a pure german language model, arXiv preprint arXiv:2012.02110 (2020).

[25] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, M. Zhou, InfoXLM: An information-theoretic framework for cross-lingual language model pre-training, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3576–3588. URL: https://aclanthology.org/2021.naacl-main.280/. doi:10.18653/v1/2021.naacl-main.280.

[26] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[27] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).