QU-NLP at CheckThat! 2025: Multilingual Subjectivity in News Articles Detection Using Feature-Augmented Transformer Models with Sequential Cross-Lingual Fine-Tuning

Mohammad AL-Smadi

Qatar University, Doha, Qatar

Abstract

This paper presents our approach to the CheckThat! 2025 Task 1 on subjectivity detection, where systems are challenged to distinguish whether a sentence from a news article expresses the subjective view of the author or presents an objective view on the covered topic. We propose a feature-augmented transformer architecture that combines contextual embeddings from pre-trained language models with statistical and linguistic features. Our system leveraged pre-trained transformers with additional lexical features: for Arabic we used AraELECTRA augmented with part-of-speech (POS) tags and TF-IDF features, while for the other languages we fine-tuned a cross-lingual DeBERTa V3 model combined with TF-IDF features through a gating mechanism. We evaluated our system in monolingual, multilingual, and zero-shot settings across multiple languages including English, Arabic, German, Italian, and several unseen languages. The results demonstrate the effectiveness of our approach, achieving competitive performance across different languages with notable success in the monolingual setting for English (rank 1st with macro-F1=0.8052), German (rank 3rd with macro-F1=0.8013), Arabic (rank 4th with macro-F1=0.5771), and Romanian (rank 1st with macro-F1=0.8126) in the zero-shot setting. We also conducted an ablation analysis that demonstrated the importance of combining TF-IDF features with the gating mechanism and the cross-lingual transfer for subjectivity detection. Furthermore, our analysis reveals the model's sensitivity to both the order of cross-lingual fine-tuning and the linguistic proximity of the training languages.

Keywords

Subjectivity Detection, Multilingual NLP, Cross-lingual Transfer, Transformer Models, AraELECTRA, DeBERTa V3, TF-IDF Features, POS Tagging, Zero-shot Learning

1. Introduction

The rapid increase in online news and social media posts has led to an crucial need for automated tools that can distinguish between factual reporting and opinion-based content. Subjectivity detection is defined as the task of identifying whether a text expresses personal opinions, beliefs, feelings, or judgments versus presenting only factual information [1]. Subjectivity detection has become a critical component in various natural language processing applications, including media bias detection, stance detection, and fact-checking services. Moreover, Subjectivity detection tools have the ability to automatically identify subjective content in multilingual contexts, whereas manual analysis is expensive and time consuming across different languages.

The CheckThat! Lab at CLEF 2025 [2] introduced Task 1 on subjectivity detection, challenging participants to develop systems capable of classifying sentences from news articles as either subjective (SUBJ) or objective (OBJ). This task was structured into three distinct settings: monolingual (training and testing in the same language), multilingual (training and testing on data comprising several languages), and zero-shot (training on several languages and testing on unseen languages). This comprehensive evaluation framework allows for a thorough assessment of systems' capabilities to generalize across languages and domains.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
malsmadi@qu.edu.qa (M. AL-Smadi)
0000-0002-7808-6962 (M. AL-Smadi)

In this paper, we present the approach developed by our team QU-NLP[1] for the CheckThat! 2025 Task 1. Our models leverage a feature-augmented transformer architecture that combines the contextual learning capabilities of pre-trained language models with statistical and linguistic features specifically selected to capture signs of subjectivity. As the task covers different language settings, we employed tailored models: (a) a specialized AraELECTRA-based model for Arabic and (b) a DeBERTa-based architecture with sequential cross-lingual fine-tuning for other languages.

Our contributions can be summarized as follows:

- We propose a feature-augmented transformer architecture that effectively combines deep contextual representations with explicit linguistic features for subjectivity detection.
- We demonstrate the effectiveness of sequential cross-lingual fine-tuning for improving performance in multilingual and zero-shot settings.
- We provide a comprehensive analysis of our system's performance across different languages and settings, highlighting strengths and limitations.
- We investigate the contribution of different feature combinations to the overall performance, offering insights into the linguistic markers of subjectivity across languages.

The remainder of this paper is organized as follows: Section 2 reviews related work in subjectivity detection and multilingual NLP. Section 3 describes the task, datasets, and our methodology, including model architecture and training setup. Section 4 presents our experimental results across different languages and settings. Section 5 discusses our findings, analyzes error cases, and explores the implications of our results. Finally, Section 6 concludes the paper and suggests directions for future work.

2. Related Work

Subjectivity detection has been an active area of research in natural language processing for over two decades. Early approaches to this task relied heavily on lexical resources and hand-crafted features [3], while more recent methods leverage deep learning architectures and transfer learning from pre-trained language models. In this section, we review relevant literature on subjectivity detection, multilingual approaches to text classification, and recent advances in cross-lingual transfer learning.

2.1. Multilingual Text Classification

Multilingual text classification has gained significant attention with the development of cross-lingual embeddings and multilingual pre-trained language models. Cross-lingual transfer learning aims to leverage knowledge from resource-rich languages to improve performance on low-resource languages. Various approaches have been proposed to enhance cross-lingual transfer, including adversarial training, meta-learning, and language-specific adapters. Artetxe and Schwenk [4] proposed a language-agnostic sentence embedding model trained on parallel data from 93 languages, enabling zero-shot cross-lingual transfer for various classification tasks.

Multilingual pre-trained language models such as mBERT [5], XLM-R [6], and mT5 [7] have become the foundation for state-of-the-art multilingual text classification systems. These models are pre-trained on massive multilingual corpora, allowing them to learn shared representations across languages that can be fine-tuned for specific downstream tasks.

Several studies have explored techniques to improve cross-lingual transfer in text classification. Wu and Dredze [8] analyzed the cross-lingual capabilities of mBERT across 39 languages and 5 NLP tasks, finding that it performs remarkably well even for languages with limited pre-training data. Pires et al. [9] investigated the structural similarities captured by mBERT that enable its cross-lingual abilities, showing that it aligns representations of similar words across languages.

[1] the team name was set to the username default value of "msmadi" on the codalab website, See tasks' final results on https://gitlab.com/checkthat_lab/clef2025-checkthat-lab/-/tree/main/task1

Yan et al. [10] proposed a meta-learning approach for cross-lingual transfer, where a model learns to quickly adapt to new languages with minimal supervision. Pfeiffer et al. [11] introduced MAD-X, a modular adaptation framework that uses language adapters to enable parameter-efficient cross-lingual transfer.

Sequential fine-tuning has emerged as an effective technique for cross-lingual transfer. Do and Gaspers [12] demonstrated that sequentially fine-tuning a multilingual model on related languages before the target language can significantly improve performance. Similarly, Nooralahzadeh et al. [13] showed that intermediate fine-tuning on a related high-resource language can boost zero-shot performance on low-resource languages.

2.2. Subjectivity Detection in News Media

Subjectivity detection in news media presents unique challenges due to the different ways in which subjective content can be expressed in seemingly objective reporting. The task of distinguishing between subjective and objective text has its roots in the pioneering work of Wiebe et al. [14], who created one of the first annotated corpora for subjectivity analysis. This early work established the foundation for subsequent research on subjectivity detection, sentiment analysis, and opinion mining. Recasens et al. [15] identified linguistic indicators of bias in news articles, including factive verbs, implicative verbs, and hedges, which can signal subjective content without explicit opinion markers.

Subjectivity detection in Arabic news has gained increasing attention over the past two decades, with researchers aiming to distinguish between factual reporting and opinionated content in Arabic-language media. Early foundational work by El-Halees [16] explored text classification in Arabic news using machine learning techniques such as maximum entropy, laying the groundwork for subsequent efforts in identifying subjective language in formal Arabic contexts. Abdul-Mageed and Diab [17] advanced the field by developing supervised models to detect subjectivity and sentiment in Modern Standard Arabic, demonstrating the viability of using linguistic features and annotated corpora for reliable classification. More recent research by Al-Smadi et al. [18] introduced an aspect-based sentiment analysis framework tailored to Arabic news articles, marking a shift from document-level to aspect-level opinion mining. While not the first to explore subjectivity in Arabic, this study is notable for its emphasis on identifying sentiment tied to specific news aspects, thereby offering a more nuanced understanding of reader affect.

Recent work has focused on developing fine-grained approaches to detect different types of subjectivity in news. Spinde et al. [19] created a comprehensive framework for detecting media bias, incorporating subjectivity detection as a key component.

The CheckThat! Lab has contributed significantly to advancing research in this area by providing multilingual benchmarks for subjectivity detection in news. The annotation guidelines developed by Ruggeri et al. [1] provide a language-agnostic framework for identifying subjectivity, enabling consistent annotation across different languages. Building on this work, Antici et al. [20] created a corpus for sentence-level subjectivity detection in English news articles, while Suwaileh et al. [21] developed ThatiAR, a dataset for subjectivity detection in Arabic news sentences.

Our work builds upon these foundations, leveraging insights from both subjectivity detection research and cross-lingual transfer learning to develop a robust system for multilingual subjectivity detection in news media.

3. Research Methodology

3.1. Task Description

The CheckThat! 2025 Task 1 focused on subjectivity in news articles detection. Participants were requested to develop systems capable of distinguishing whether a sentence from a news article expresses the subjective view of the author or presents an objective view on the covered topic. This binary classification task required systems to label text sequences as either subjective (SUBJ) or objective (OBJ).

The task was structured into three distinct evaluation settings:

Table 1
Dataset Statistics for CheckThat! Task 1 - Subjectivity

Language	Train		Dev		Dev-Test	
	OBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ
English	532	298	222	240	362	122
Italian	1231	382	490	177	377	136
German	492	308	317	174	226	111
Arabic	1391	1055	266	201	425	323

1. **Monolingual**: Systems were trained and tested on data in a single language. This setting was implemented for five languages: English, Arabic, Italian, and German.
2. **Multilingual**: Systems were trained and tested on data from several languages.
3. **Zero-shot**: Systems were trained on several languages from the settings above and tested on unseen languages (mainly Polish, Ukrainian, Romanian, and Greek).

The participating systems were ranked based on their macro-averaged F1 score, which equally weights the performance on both the SUBJ and OBJ classes.

3.2. Dataset

The dataset provided for the task consisted of sentences extracted from news articles in multiple languages, manually annotated as either subjective (SUBJ) or objective (OBJ). For each language, the data was divided into three sets: training, development, and test.

Table 1 presents the statistics of the dataset for each language. The data exhibits some class imbalance, with objective sentences generally outnumbering subjective ones across most languages. This imbalance varies across languages, with Arabic having the largest dataset (3,661 annotated sentences) and German having the smallest (1,628 annotated sentences). About 300 sentences were provided as test dataset for each language.

The annotation of the dataset followed the guidelines developed by Ruggeri et al. [1], which provide a language-agnostic framework for identifying subjectivity in news text. These guidelines define subjective content as text that expresses personal opinions, beliefs, or judgments, while objective content presents factual information without expressing the author's perspective. The reader is redirected to [20, 21, 2] for more information about the datasets.

3.3. Models

Our approach to the subjectivity detection task involved developing two distinct model architectures tailored to different language settings. For the Arabic monolingual task, we designed a specialized model leveraging AraELECTRA with additional linguistic features. For all other settings (monolingual non-Arabic, multilingual, and zero-shot), we employed a DeBERTa-based architecture with sequential cross-lingual fine-tuning. The upcoming sub-sections explain in more detail the models architectures along with thier training setups.

3.3.1. Arabic Monolingual Model

We developed a feature-augmented transformer architecture for Arabic, leveraging the AraELECTRA model [22]. This architecture integrates the pre-trained language model's contextual understanding with supplementary linguistic features. Specifically, it incorporates Part-of-Speech (POS) tags and Term Frequency-Inverse Document Frequency (TF-IDF) representations to capture subjectivity markers in Arabic text.

The proposed model builds upon ELECTRA [23] and its Arabic adaptation, AraELECTRA [22]. ELECTRA is an encoder-only transformer designed for enhanced efficiency in Natural Language

Processing (NLP) tasks. Unlike traditional Masked Language Models (MLMs), ELECTRA employs a "replaced token detection" training strategy. While models like BERT [24] predict masked words, ELECTRA's generator component proposes plausible alternative tokens. A discriminator then identifies whether each input token is original or replaced. This unique strategy compels the model to learn from all input tokens, rather than just masked ones. Consequently, this approach boosts model efficiency and reduces the required training epochs.

The model consists of the following components:

1. **Backbone Encoder**: We used the pre-trained `araelectra-base-discriminator`[2] as the core of our model [22]. The [CLS] token from the final hidden layer is passed through a self-attention module (MultiheadAttention) to obtain a refined representation.
2. **Part-of-Speech Features**: We extracted POS tag distributions using the `bert-base-arabic-camelbert-mix-pos-msa`[3] model [25]. The resulting 9-dimensional POS tag distribution is projected to 64 dimensions via a linear layer followed by Rectified Linear Unit (ReLU) activation function. Applied after linear layers, ReLU enables models to learn complex data patterns [26]. This boosts the model's ability to recognize deep and complex relationships.
3. **TF-IDF Features**: We computed TF-IDF features over character n-grams (3-7) using a TfidfVectorizer. The resulting vector is reduced to 128 dimensions through a learnable projection layer with ReLU activation.
4. **Fusion and Classification**: The refined [CLS] embedding from AraELECTRA (768 dimensions), the POS projection (64 dimensions), and the TF-IDF projection (128 dimensions) are concatenated into a 960-dimensional feature vector. This vector is then passed through a fully connected network consisting of a linear layer ($960 \rightarrow 512$) followed by LayerNorm and Dropout, and a final linear layer ($512 \rightarrow 2$) for binary classification.

3.3.2. DeBERTa-based Model for Other Languages

For non-Arabic languages and the multilingual/zero-shot settings, we developed a model based on the DeBERTa V3 architecture [27] with a gating mechanism for integrating lexical features. This model was designed to effectively transfer knowledge across languages through sequential fine-tuning.

The model architecture includes:

1. **DeBERTa V3 Encoder**: We used the `deberta-v3-large`[4] model as our backbone. The encoder outputs are passed through a 16-head self-attention layer to capture richer inter-token dependencies. We extract the representation corresponding to the [CLS] token and apply layer normalization and dropout.
2. **TF-IDF Lexical Branch**: We extract lexical features using a TfidfVectorizer with character n-grams (3-7). The resulting sparse matrix is projected into a dense 128-dimensional vector via a feedforward layer.
3. **Gating Mechanism**: A gating scalar is computed to dynamically weigh the importance of lexical versus contextual information. This gate modulates the 128-dimensional TF-IDF embedding.
4. **Feature Fusion and Classification**: The gated TF-IDF vector and the DeBERTa-derived CLS embedding are concatenated and passed through a classification head consisting of linear layers, layer normalization, ReLU activation, and dropout.

3.4. Gating Mechanism for Feature Fusion

To effectively integrate sparse lexical representations with dense contextual embeddings, our model employs a learnable *gating mechanism* that dynamically modulates the contribution of TF-IDF features based on the semantic richness of the input as captured by DeBERTaV3.

[2] https://huggingface.co/aubmindlab/araelectra-base-discriminator
[3] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa
[4] https://huggingface.co/microsoft/deberta-v3-large

Let $\mathbf{h}_{\text{BERT}} \in \mathbb{R}^d$ denote the contextualized representation derived from the [CLS] token output of the DeBERTaV3 encoder, where d is the hidden size of the transformer. The TF-IDF vector, denoted $\mathbf{h}_{\text{TFIDF}} \in \mathbb{R}^k$, is passed through a fully connected layer with ReLU activation to yield $\tilde{\mathbf{h}}_{\text{TFIDF}} \in \mathbb{R}^{128}$, enhancing its representational capacity. The gating mechanism then computes a scalar gate value:

$$g = \sigma(\mathbf{W}_g \mathbf{h}_{\text{BERT}} + b_g)$$

where $\mathbf{W}_g \in \mathbb{R}^{1 \times d}$, $b_g \in \mathbb{R}$, and $\sigma(\cdot)$ is the sigmoid activation function. This scalar gate $g \in [0, 1]$ acts as a dynamic weighting coefficient:

$$\hat{\mathbf{h}}_{\text{TFIDF}} = g \cdot \tilde{\mathbf{h}}_{\text{TFIDF}}$$

The modulated TF-IDF vector $\hat{\mathbf{h}}_{\text{TFIDF}}$ is then concatenated with $\mathbf{h}_{\text{BERT}}$ to form the joint representation:

$$\mathbf{h}_{\text{joint}} = [\mathbf{h}_{\text{BERT}}; \hat{\mathbf{h}}_{\text{TFIDF}}]$$

This joint vector is subsequently passed through a feedforward layer followed by a classifier to produce the final output logits.

The gating mechanism enables the model to adaptively regulate the influence of TF-IDF features on a per-instance basis. When semantic signals from the pretrained language model are strong, the gate may downscale the TF-IDF contribution. Conversely, in scenarios where domain-specific vocabulary or sparse lexical cues offer additional value, the gate enhances their impact. This dynamic fusion strategy improves robustness across domains and languages by learning to balance deep semantic understanding with interpretable lexical signals. This approach draws inspiration from prior work on *Highway Networks* [28] and feature gating mechanisms in multimodal learning, where learned gates enable networks to dynamically fuse heterogeneous input modalities.

3.5. Training Setup

3.5.1. Arabic Monolingual Model Training

For the Arabic model, we employed the following training configuration:

- **Preprocessing**: Input text was tokenized using the ELECTRA tokenizer with a maximum length of 512 tokens. POS tag distributions were normalized, and TF-IDF vectors were computed with a maximum of 3000 features and a minimum document frequency of 2.
- **Training Parameters**: We used a learning rate of 1e-5, a batch size of 16, and gradient accumulation of 4 steps. The model was trained for up to 100 epochs with early stopping (patience = 3) based on evaluation loss. We applied weight decay of 0.01 and enabled mixed precision training (fp16) for efficiency.
- **Optimization**: We used the AdamW optimizer with a linear learning rate scheduler and 100 warmup steps.
- **Evaluation**: The model was evaluated after each epoch using the development set, and the best checkpoint was selected based on the lowest evaluation loss.

3.5.2. DeBERTa-based Model Training

For the DeBERTa-based model, we implemented a sequential cross-lingual fine-tuning approach:

- **Preprocessing**: Sentences were tokenized using the DeBERTaV2Tokenizer with a maximum length of 512 tokens. TF-IDF features were extracted from the training data and saved for later use.
- **Sequential Fine-tuning**: We trained the model in a specific language sequence: [German $\rightarrow$ Italian $\rightarrow$ English]. Starting with the base `microsoft/deberta-v3-large` checkpoint, we fine-tuned on German data, then used the resulting model to fine-tune on Italian data, and finally fine-tuned on English data.

Table 2
Monolingual Results (Macro F1 scores)

Language	QU-NLP (Our Team)	Top Team	Baseline
English	0.8052 (1st)	0.8052 (QU-NLP)	0.5370
German	0.8013 (3rd)	0.8520 (smollab)	0.6960
Italian	0.7139 (7th)	0.8104 (XplaiNLP)	0.6941
Arabic	0.5771 (4th)	0.6884 (CEA-LIST)	0.5133

Table 3
Multilingual Results (Macro F1 scores)

Setting	QU-NLP (Our Team)	Top Team	Baseline
Multilingual	0.6692 (8th)	0.7550 (TIFIN India)	0.6390

- **Training Parameters**: We used a learning rate of 1e-5, a batch size of 8, and gradient accumulation of 2 steps. Each language-specific fine-tuning was run for up to 100 epochs with early stopping (patience = 2) based on evaluation loss. We applied weight decay of 0.01 and used a cosine learning rate scheduler with 100 warmup steps.
- **Multilingual and Zero-shot Setting**: For both multilingual and zero-shot evaluation, we evaluated the model fine-tuned on the sequence of languages (German $\rightarrow$ Italian $\rightarrow$ English) without any additional training on the target languages.

4. Results

In this section, we present the results of our systems across the three evaluation settings: monolingual, multilingual, and zero-shot. We compare our performance with other participating teams and analyze the effectiveness of our approaches for different languages. For more information about the baseline models or the other participating teams' models the reader is redirected to [2].

4.1. Monolingual Results

Table 2 presents the results of our system in the monolingual setting for each of the four languages, along with the performance of the top-performing team and the baseline for each language.

Our system achieved the best performance for English with a macro F1 score of 0.8052, significantly outperforming the baseline (0.5370). For German, our system ranked third with a macro F1 score of 0.8013, competitive with the top-performing team (0.8520). In Italian, our system achieved a macro F1 score of 0.7139, ranking seventh but still substantially above the baseline (0.6941). For Arabic, our system ranked fourth with a macro F1 score of 0.5771, which is above the baseline (0.5133) but considerably lower than the top-performing team (0.6884).

4.2. Multilingual Results

Table 3 shows the results of our system in the multilingual setting, where DeBERTa-based models were trained on data from the monolingual setting and evaluated on the multilingual test data.

In the multilingual setting, our system achieved a macro F1 score of 0.6692, ranking eighth among all participating teams. While this performance is above the baseline (0.6390), it is notably lower than our monolingual results for English and German. This suggests that the multilingual model faces challenges in effectively learning shared representations across languages, possibly due to linguistic differences or imbalances in the training data. A clear limitation in our used DeBERTa-based model is coming from the inclusion of sentences in Arabic as part of the testing dataset, whereas the model was not trained on the Arabic language as part of the cross-lingual sequence training explained earlier.

Table 4
Zero-shot Results (Macro F1 scores)

Language	QU-NLP (Our Team)	Top Team	Baseline
Romanian	0.8126 (1st)	0.8126 (QU-NLP)	0.6461
Polish	0.5165 (13th)	0.6922 (CEA-LIST)	0.5719
Ukrainian	0.6168 (8th)	0.6424 (CSECU-Learners)	0.6296
Greek	0.4057 (11th)	0.5067 (AI Wizards)	0.4159

4.3. Zero-shot Results

Table 4 presents the results of our system in the zero-shot setting, where models were evaluated on languages not seen during training.

Our system demonstrated varying performance across the zero-shot languages. For Romanian, we achieved the best performance among all teams with a macro-F1 score of (0.8126), significantly outperforming the baseline (0.6461). This suggests that our sequential fine-tuning approach effectively transferred knowledge to Romanian, possibly due to linguistic similarities with the training languages.

However, for Polish, Ukrainian, and Greek, our system's performance was less impressive. In Polish, we ranked 13th with a macro-F1 score of (0.5165), which is below the baseline (0.5719). In Ukrainian, we ranked 8th with a score of 0.6168, slightly below the baseline (0.6296). In Greek, we ranked 11th with a score of 0.4057, slightly below the baseline (0.4159).

5. Discussion

Our participation in the CheckThat! 2025 Task 1 on subjectivity detection yielded several insights into the effectiveness of different approaches for this task across languages and evaluation settings. In this section, we discuss our findings, analyze the strengths and limitations of our approach, and explore potential avenues for improvement.

5.1. Analysis of Model Performance

The performance of our systems varied considerably across languages and evaluation settings, revealing several interesting patterns:

- **Strong Monolingual Performance**: Our models performed particularly well in the monolingual setting for English and German, achieving F1 scores of 0.8052 and 0.8013, respectively. This suggests that our feature-augmented transformer architecture effectively captures markers of subjectivity in these languages.
- **Varying Cross-lingual Transfer**: The effectiveness of cross-lingual learning transfer varied significantly across target languages. The outstanding performance on Romanian (F1=0.8126) in the zero-shot setting demonstrates that our sequential fine-tuning approach can successfully transfer knowledge to linguistically similar languages. However, the relatively poor performance on Polish, Ukrainian, and Greek suggests limitations in transferring to more distant languages.
- **Multilingual vs. Monolingual Trade-off**: Our multilingual model (F1=0.6692) underperformed compared to our best monolingual models, highlighting the challenges of developing a single model that performs well across multiple languages simultaneously. In addition, the Arabic language was not included as part of the cross-lingual sequence training of the model evaluated using the multilingual dataset.

5.2. Feature Contribution Analysis

To understand the contribution of different features to our DeBERTa-based model's performance, we conducted an ablation study on the English, German and Italian monolingual models. Table 6 presents

Table 5
Ablation Study on German, Italian, and English, monolingual models with the sequence of [German → Italian → English] during training

Model Configuration	Macro F1 (German)	Macro F1 (Italian)	Macro F1 (English)
Full Model (DeBERTa + TF-IDF + Gating)	0.8013	0.7139	0.8052
DeBERTa Only	0.5866	0.7040	0.7974
DeBERTa + TF-IDF (No Gating)	0.5245	0.7234	0.7707
Full Model (without cross-lingual training)	0.8013	0.6920	0.7818

Table 6
F1-macro results for the English, German and Italian Monolingual Models based on language order during training

Language Order	Macro F1 (German)	Macro F1 (Italian)	Macro F1 (English)
(German → Italian → English)	0.8013	0.7139	0.8052
(English → Italian → German)	0.8195	0.7787	0.7818
(German → English → Italian)	0.8013	0.8033	0.7839

the results of this analysis. Trainings of the monolingual languages forllowed the same sequence of languages [German → Italian → English] as performed in the full models' trainings.

The ablation study reveals that each component of our model contributes to its overall performance:

- The base DeBERTa model alone achieved a respectable macro-F1 score of (0.5866, 0.7974, 0.7040) for German, Italian, and English consequently, demonstrating the strong foundation provided by the pre-trained language model when combined with with cross-lingual sequence training.
- Adding TF-IDF features without the gating mechanism improved performance to (0.7234) for Italian language only, indicating that lexical features do not provide complementary information to the contextual embeddings for all languages.
- The gating mechanism further improved performance, allowing the model to dynamically balance the contribution of lexical features complementing information gained from the contextual embeddings for English and German languages.
- The full model, combining DeBERTa, TF-IDF features, and the gating mechanism, achieved the best performance of (0.8052, 0.8013) for English and German consequently, confirming the value of our feature-augmented approach.
- The cross-lingual sequence training positively enhanced the monolingual models' results, where training the full monolingual DeBERTa-based model without the cross-lingual sequence training achieved lower results of (0.7818, 0.6920) compared to (0.8052, 0.7139) for English and Italian languages consequently.

5.3. Language Order in Cross-lingual Training

The results of our cross-lingual subjectivity detection experiments demonstrate a notable sensitivity to the ordering of language fine-tuning. The results presented in Table 6 demonstrate that the ordering of languages during cross-lingual fine-tuning has a significant impact on model performance across English, German, and Italian. Notably, the model trained in the sequence [*English* → *Italian* → *German*] achieves the highest F1 score on German (0.8195), and a strong improvement on Italian (0.7787), albeit with a slight drop in English performance (0.7818). In contrast, the sequence [*German* → *Italian* → *English*] results in the lowest Italian score (0.7139), while preserving high performance on German (0.8013) and English (0.8052). Interestingly, training in the order [*German* → *English* → *Italian*] yields the best Italian performance (0.8033), suggesting a complex interaction between intermediate representations and language-specific features.

These results provide several insights into the dynamics of multilingual transfer for subjectivity detection. First, the finding that German performance improves when preceded by English suggests that English provides beneficial representations which transfer well to German, a typologically related language [6, 29]. This supports prior work showing that English often acts as a strong base model for multilingual tasks due to its central position in pretrained multilingual language models.

Second, the decline in Italian performance when preceded by German (0.7139) compared to when preceded by English (0.7787) or English and German (0.8033) is indicative of language interference and potential catastrophic forgetting [30, 31]. German's syntactically rigid and morphologically rich characteristics may interfere with learning semantic cues for Italian, a Romance language that relies more heavily on pragmatic and lexical signals of subjectivity [9].

Third, the sequence [*German* $\rightarrow$ *English* $\rightarrow$ *Italian*] achieving the highest Italian F1 score implies a positive cumulative effect when both typologically diverse languages precede Italian. This ordering may allow the model to retain robust representations for both syntactic (from German) and semantic-pragmatic (from English) subjectivity features before learning Italian, thereby enabling better generalization.

Lastly, the variations in English scores across the setups (ranging from 0.7818 to 0.8052) suggest that English benefits from being either the final or intermediate fine-tuning target but may degrade when trained first—likely due to subsequent overwriting of its learned representations. This aligns with recent findings on cross-lingual anchoring effects, where the initial language in training can disproportionately shape the shared representational space [32].

6. Conclusion

In this paper, we presented our approach to the CheckThat! 2025 Task 1 on subjectivity detection, which challenged participants to distinguish between subjective and objective sentences in news articles across multiple languages. Our system leveraged feature-augmented transformer architectures, combining the contextual understanding capabilities of pre-trained language models with statistical and linguistic features specifically designed to capture markers of subjectivity.

Results demonstrated the effectiveness of our approach, particularly in the monolingual setting for English and German, and in the zero-shot setting for Romanian. The strong performance on Romanian highlights the potential of our sequential cross-lingual fine-tuning approach for transferring knowledge to linguistically similar languages. However, the varying performance across languages and evaluation settings also revealed challenges in developing truly language-agnostic models for subjectivity detection. The ablation study confirmed the value of our feature-augmented approach, showing that each component of our model contributed to its overall performance.

Our findings reinforce the importance of language order in cross-language fine-tuning and suggest linguistic proximity (i.e. how similar two languages are to each other), and task-specific signal transfer (i.e. how well the model can recognize and use these specific opinion-indicating cues from one language when trying to understand opinions in another), should all be considered when designing cross-lingual pipelines for subjectivity detection. For instance, a language might use certain common phrases or word endings to show an opinion, while another might rely more on the speaker's tone or context.

Future work could explore several promising directions for improving multilingual subjectivity detection. These include developing more sophisticated language-specific features, implementing adversarial training techniques to create more language-agnostic representations, generating synthetic training data to address class imbalance, exploring multi-task learning approaches, and developing ensemble methods that combine the strengths of multiple specialized models.

Declaration on Generative AI

During the preparation of this work, the author(s) used Overleaf Writefull service in order to: Grammar and spelling check. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] F. Ruggeri, F. Antici, A. Galassi, A. Korre, A. Muti, A. Barron, On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection, Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts 3370 (2023) 103–111.

[2] 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.

[3] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, Learning subjective language, Computational Linguistics 30 (2004) 277–308.

[4] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (2019) 4171–4186.

[6] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[7] 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, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021) 483–498.

[8] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (2019) 833–844.

[9] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493/. doi:10.18653/v1/P19-1493.

[10] M. Yan, H. Zhang, D. Jin, J. T. Zhou, Multi-source meta transfer for low resource multiple-choice question answering, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7331–7341.

[11] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, Mad-x: An adapter-based framework for multi-task cross-lingual transfer, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020) 7654–7673.

[12] Q. Do, J. Gaspers, Cross-lingual transfer learning with data selection for large-scale spoken language understanding, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 1455–1460.

[13] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, I. Augenstein, Zero-shot cross-lingual transfer with meta learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4547–4562.

[14] 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.

[15] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky, Linguistic models for analyzing and detect-

ing biased language, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 1 (2013) 1650–1659.
[16] A. M. El-Halees, Arabic text classification using maximum entropy, in: The International Arab Conference on Information Technology (ACIT), 2011.
[17] M. Abdul-Mageed, M. Diab, Subjectivity and sentiment analysis of modern standard arabic, in: Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL, 2011, pp. 35–44. URL: https://aclanthology.org/W11-1703/.
[18] M. Al-Smadi, M. Al-Ayyoub, H. Al-Sarhan, Y. Jararweh, An aspect-based sentiment analysis approach to evaluating arabic news affect on readers, Journal of Universal Computer Science 22 (2016) 630–649.
[19] T. Spinde, L. Rudnitckaia, J. Mitrović, F. Hamborg, M. Granitzer, B. Gipp, K. Donnay, Automated identification of bias inducing words in news articles using linguistic and context-oriented features, Information Processing & Management 58 (2021) 102505.
[20] F. Antici, F. Ruggeri, A. Galassi, A. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, et al., 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), ELRA and ICCL, 2024, pp. 273–285.
[21] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, Thatiar: Subjectivity detection in arabic news sentences, arXiv preprint arXiv:2406.05559 (2024).
[22] W. Antoun, F. Baly, H. Hajj, Araelectra: Pre-training text discriminators for arabic language understanding, arXiv preprint arXiv:2012.15516 (2020).
[23] K. Clark, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).
[24] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, Minneapolis, Minnesota, 2019, p. 2.
[25] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, N. Habash, The interplay of variant, size, and task type in Arabic pre-trained language models, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Online), 2021.
[26] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
[27] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.
[28] R. K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, 2015. URL: https://arxiv.org/abs/1505.00387. arXiv:1505.00387.
[29] E. P. Stabler, E. L. Keenan, Structural similarity within and among languages, Theoretical Computer Science 293 (2003) 345–363.
[30] Z. Li, D. Hoiem, Learning without forgetting, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 2935–2947.
[31] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences 114 (2017) 3521–3526.
[32] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, et al., Crosslingual generalization through multitask finetuning, in: The 61st Annual Meeting Of The Association For Computational Linguistics, 2023.