TIFIN at CheckThat! 2025: Cross-Lingual Subjectivity Classification in News through Monolingual, Multilingual, and Zero-Shot Learning*

Notebook for the CheckThat! Lab at CLEF 2025

Kishan Gurumurthy[1,*,†], Ashish Shrivastava[1,*,†], Pawan Kumar Rajpoot[1], Prasanna Devadiga[1], Bharatdeep Hazarika[1], Manish Jain[1], Manan Sharma[1], Arya Suneesh[1], Anshuman B Suresh[1] and Aditya U Baliga[1]

[1]*TIFIN India*

Abstract

In an age of widespread digital misinformation, the process of binary classification to differentiate subjective claims from objective reporting is crucial for building efficient automated fact-checking systems. This paper presents our approach for Task 1 of the CLEF 2025 CheckThat! Lab, which requires classifying text segments as either subjective or objective. The evaluation spans three settings—monolingual, multilingual, and zero-shot cross-lingual transfer—across five languages: Arabic, Bulgarian, English, German, and Italian. Our method leverages pretrained transformer-based language models that are fine-tuned specifically for subjectivity detection, with adaptations designed to enhance performance in multilingual and cross-lingual contexts. To address the issue of class imbalance present in the training data, we incorporate resampling and class-weighting techniques during model training, which significantly improve the identification of less frequent classes. Experimental results show consistent and strong performance across all evaluation settings, particularly in scenarios involving limited resources and unseen languages. Additionally, comprehensive error analysis is conducted to explore linguistic and contextual influences on classification accuracy. These results demonstrate the importance of robust multilingual modeling approaches in subjectivity detection and their contribution to advancing automated fact-checking and the promotion of reliable information dissemination.

Keywords

subjectivity classification, fact-checking automation, multilingual modeling, cross-lingual generalization, class imbalance mitigation

1. Introduction

In computational linguistics, the distinction between subjective and objective language plays a pivotal role in various natural language processing (NLP) tasks and applications [1, 2]. Subjectivity refers to the personal opinions, beliefs, and emotions, while objectivity denotes factual reporting devoid of personal bias or interpretation [3, 4]. The subjectivity detection task, in the context of news articles, is a binary classification task that has garnered significant attention in recent years [5, 6]. This paper discusses our participation in the CLEF 2025 Task 1: Subjectivity Detection, a competition aimed at advancing methodologies for distinguishing between subjective (SUBJ) and objective (OBJ) sentences across three distinct settings: monolingual, multilingual, and zero-shot cross-lingual transfer.

The delineation between subjective and objective language is not merely a linguistic exercise; it is a fundamental challenge in NLP that has far-reaching implications for how information is processed and understood [7, 8]. Subjective sentences often contain evaluative language, emotional undertones,

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

*Corresponding author.

†These authors contributed equally to this work.

Kishan.gurumurthy@workifi.com (K. Gurumurthy); Ashish.shrivastava@workifi.com (A. Shrivastava); pawan@tifin.com (P. K. Rajpoot); prasanna@askmyfi.com (P. Devadiga); bharatdeep@askmyfi.com (B. Hazarika); manish.jain@tifin.com (M. Jain); manan.sharma@tifin.com (M. Sharma); arya.suneesh@tifin.com (A. Suneesh); anshuman.suresh@tifin.com (A. B. Suresh); aditya@askmyfi.com (A. U. Baliga)

and personal perspectives, making them inherently ambiguous and context-dependent [9, 10]. For instance, a statement like "The film was thrilling" embodies a subjective viewpoint, colored by the speaker's personal experience and emotions. In contrast, an objective sentence such as "The film was released in 2023" presents verifiable information devoid of personal sentiment [11]. Automated subjectivity detection is crucial for various reasons [12, 13]. First, it serves as a foundational step in sentiment analysis, where distinguishing between subjective and objective content is essential to accurately assess public sentiment [14, 15, 16]. Furthermore, the inherent ambiguity of subjective language poses challenges for computational models, which must be equipped to navigate the complexities of context and nuance [17, 18]. This complexity is exacerbated in multilingual contexts, where cultural variations influence how subjectivity is expressed [19, 20]. Different languages may employ distinct syntactic structures, lexical choices, and idiomatic expressions to convey subjective nuances, thus complicating the task of developing universally applicable models [21, 22].

The implications of effective subjectivity detection extend to multiple domains, particularly in media analysis and information retrieval [23, 24]. In news media analysis, the ability to distinguish between factual reporting and opinion journalism is vital for maintaining journalistic integrity and informing readers accurately [25, 26]. For example, a news article that presents a politician's statement as fact without context may mislead the audience; thus, identifying subjective content is crucial for responsible reporting [27, 28]. In the realm of social media monitoring, subjectivity detection enables platforms to better understand user sentiment and identify potentially biased or misleading content [29, 30]. This capability is particularly important for combating the spread of misinformation and propaganda, especially during critical events such as elections or public health crises [31, 32]. Businesses have also found subjectivity detection incredibly valuable for understanding what customers really think about their products and services. When companies analyze online reviews and social media posts, being able to separate factual complaints from emotional reactions helps them make better decisions about product improvements and marketing strategies[33, 34]. Information retrieval systems also benefit significantly from subjectivity detection [35, 36]. Search engines and information retrieval systems get a similar boost from this technology - imagine how much more useful search results would be if they could automatically flag whether a piece of content is presenting facts or someone's personal opinion. For instance, a user searching for factual information about a medical condition should receive objective, evidence-based content rather than subjective personal experiences [37, 38]. Furthermore, subjectivity detection is essential for automated fact-checking systems, which must differentiate between verifiable claims and opinion statements. This distinction is crucial for maintaining the accuracy and reliability of automated content verification tools [39, 40, 41].

In this paper, we outline our approach to the task of subjectivity detection within the CLEF 2025 framework. Our methodology involves leveraging advanced machine learning techniques to classify sentences as either subjective or objective across monolingual, multilingual, and zero-shot settings. Preliminary findings indicate promising performance across these various contexts, demonstrating the potential of our approach to address the challenges associated with subjectivity detection. Through this work, we hope to push forward our understanding of how machines can better distinguish between objective reporting and subjective opinion. This research matters because getting subjectivity detection right has real-world impact - from helping journalists maintain editorial standards to improving how search engines filter information, and making sentiment analysis tools more reliable. The challenge of automatically identifying subjective language remains one of the more fascinating problems in computational linguistics. As our digital world becomes increasingly saturated with opinions, personal viewpoints, and biased content, the ability to separate facts from opinions becomes not just academically interesting, but practically essential for building trustworthy NLP systems. As we navigate the complexities of subjective language, our participation in CLEF 2025 Task 1 serves as a valuable opportunity to contribute to the development of methodologies that can effectively address these challenges across diverse linguistic and cultural contexts.

2. Related Work

2.1. Foundational Work on Subjectivity Detection

Subjectivity detection has its roots in the early 2000s, with seminal works laying the groundwork for subsequent research. Wiebe et al. (1999) pioneered the field with their classification of subjective and objective sentences, introducing a lexicon-based approach that distinguished between factual and opinionated content [42]. This foundational work was expanded in Wiebe et al. (2004), where the authors elaborated on the significance of subjectivity in natural language processing (NLP) and proposed a more nuanced framework for identifying subjective expressions [43]. Pang and Lee (2004) further advanced the field by differentiating between subjectivity and sentiment analysis, emphasizing the importance of context in understanding subjective content [44]. Their later work (2008) highlighted the challenges of classifying subjective sentences within various domains, establishing a benchmark for subsequent studies [45]. Wilson et al. (2005) contributed to fine-grained opinion recognition, introducing methods to detect and classify opinions within text, thereby enhancing the granularity of subjectivity detection [4]. Yu and Hatzivassiloglou (2003) provided a critical perspective by focusing on the separation of facts from opinions, which remains a central challenge in subjectivity detection [8].

2.2. Machine Learning Approaches

2.2.1. Traditional ML Methods

The evolution of subjectivity detection has been significantly influenced by machine learning methodologies. Early approaches predominantly relied on feature-based models that utilized lexical, syntactic, and semantic features to classify sentences. Support Vector Machines (SVM) and Naive Bayes classifiers emerged as popular choices, demonstrating effective performance in various datasets. For instance, the work by Read (2005) [46] achieved an accuracy of 80% using Naive Bayes, while SVMs, as demonstrated by Zhang and Liu (2011) [47], showed superior performance with an F1 score of 0.82.

2.2.2. Deep Learning Era

The advent of deep learning marked a paradigm shift in subjectivity detection. Neural network architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been employed to capture complex patterns in textual data. Kim (2014) showcased the efficacy of CNNs in sentiment analysis, achieving state-of-the-art results on benchmark datasets [48]. The introduction of attention mechanisms and transformers has further revolutionized the field. The BERT model (Devlin et al., 2018) [49] and its variants have set new benchmarks in various NLP tasks, including subjectivity detection. Liu et al. (2019) demonstrated that fine-tuning BERT for subjectivity detection yielded significant improvements, achieving an accuracy of 92% on standard datasets [50].

2.2.3. Multilingual and Cross-lingual Approaches

As the demand for multilingual applications grew, researchers began exploring cross-lingual subjectivity detection. Cross-lingual word embeddings, such as those proposed by Mikolov et al. (2013) in their seminal Word2Vec work, facilitated the transfer of knowledge across languages [51, 52]. The development of multilingual BERT (Devlin et al., 2018) and XLM models (Lample & Conneau, 2019) has further advanced this area, allowing for effective subjectivity detection in multiple languages without the need for extensive retraining [49, 53]. Conneau et al. (2020) introduced XLM-R, which significantly outperformed multilingual BERT on cross-lingual benchmarks, demonstrating the effectiveness of scaling multilingual models with larger datasets [54].

2.3. CLEF CheckThat! Lab History

The CLEF CheckThat! Lab has played a pivotal role in advancing subjectivity detection methodologies through its annual competitions. The lab, which began in 2018, has consistently focused on fact-checking and related tasks including subjectivity detection [55]. In previous years, from 2018 to 2024, the competition has seen a variety of innovative approaches. For instance, the top-performing systems in 2019 utilized ensemble methods, combining multiple classifiers to enhance performance [56]. The 2020 competition introduced new evaluation metrics that focused on precision and recall, with the best-performing system achieving an F1 score of 0.89 [57]. The evolution of datasets and evaluation methodologies has also been noteworthy. The 2021 competition emphasized multilingual performance, with participants reporting enhanced accuracy in detecting subjectivity across diverse languages [58]. The 2022 and 2023 competitions further refined evaluation frameworks, allowing for a more comprehensive assessment of cross-lingual capabilities [59, 60]. The 2023 edition introduced multimodal approaches, with the winning system by Frick & Vogel (2023) achieving an F1 score of 0.7297 by combining textual and visual features[61].

2.4. Datasets and Resources

2.4.1. English Datasets

The construction of datasets has been crucial for the development of subjectivity detection systems. One of the most significant contributions is the "Corpus for Sentence-Level Subjectivity Detection on English News Articles," which provides a comprehensive collection of annotated sentences, facilitating the training and evaluation of models [62]. The annotation guidelines emphasize inter-annotator agreement, which has been shown to exceed 85%, underscoring the reliability of the dataset. The MPQA corpus, another foundational resource, has evolved over the years, providing rich annotations for opinionated language in news articles [4]. OpinionFinder has also been instrumental in providing tools and resources for subjectivity detection, further enriching the landscape of English datasets [63].

2.4.2. Multilingual Datasets

The need for multilingual subjectivity detection has led to the creation of cross-lingual subjectivity corpora. Resources such as the Multilingual Subjectivity Lexicon and the Multilingual Opinion Corpus (MOC) have facilitated research across various languages, including Spanish, French, German, Chinese, and Arabic [64]. These datasets have highlighted the annotation challenges posed by cultural differences in subjectivity perception [23]. The evolution of CLEF task datasets has also contributed significantly to multilingual research, providing a platform for testing and comparing methodologies across languages. Recent competitions have focused on addressing the challenges of low-resource languages, with participants developing innovative solutions to enhance subjectivity detection in these contexts.

2.5. Evaluation Methodologies

Evaluation methodologies in subjectivity detection have evolved to address the complexities of multilingual and cross-lingual settings. Standard metrics such as accuracy, F1 score, precision, and recall remain central to performance evaluation. However, the challenges of cross-lingual evaluation have necessitated the development of specialized protocols to ensure comparability across languages [65]. Recent advances in evaluation frameworks have introduced measures that account for cultural bias and domain adaptation challenges, which are critical for the accurate assessment of subjectivity detection systems [66]. The integration of these advanced methodologies has enabled researchers to better understand the strengths and weaknesses of their models in diverse linguistic contexts.

Overall, the literature on subjectivity detection has evolved significantly over the years, with foundational works paving the way for sophisticated machine learning and deep learning approaches. The CLEF CheckThat! Lab has been instrumental in driving research forward, while the development of

diverse datasets and evaluation methodologies has enriched the field. However, ongoing challenges, particularly in multilingual and low-resource contexts, highlight the need for continued research and innovation.

3. Our Approach

3.1. Data Pre-processing

We adopted different data pre-processing and enhancement strategies tailored to the three experimental settings explored in this study: (1) monolingual training and testing, (2) multilingual learning, and (3) zero-shot generalization. These configurations enabled systematic evaluation of model performance under controlled language-specific, cross-lingual, and transfer learning scenarios, particularly within the context of subjective versus objective classification.

3.1.1. Monolingual Setting

For the monolingual training and evaluation setting, we began by parsing the full development training data and isolating samples belonging exclusively to the target language under investigation. Each language was treated independently to assess the classification performance in a controlled monolingual context. Following this filtration, we conducted a statistical analysis of the class distribution across the subjective (SUBJ) and objective (OBJ) labels. Table 1 presents the class-wise distribution for each of the five target languages. A notable degree of class imbalance was observed, with the objective class typically dominating in most languages—especially in Italian and English. To mitigate this imbalance and promote better model generalization, we employed synthetic data augmentation. Specifically, we leveraged GPT-4o to generate additional examples for the underrepresented class in each language. This augmentation was performed conditionally based on the observed class distribution and was constrained to maintain semantic and syntactic coherence with the original samples. All preprocessing operations, including filtration, tokenization, and augmentation, were standardized across languages to ensure consistency and reproducibility in the experimental pipeline.

3.1.2. Multilingual Setting

For the multilingual setting, we retained the full cross-lingual dataset encompassing all languages involved in the task. A comprehensive statistical analysis was conducted to quantify the distribution of data across language partitions and subjectivity classes (SUBJ vs. OBJ). Table 1 presents the resulting class-wise distribution in both the multilingual development and training sets. This revealed significant class imbalance, especially in underrepresented language subsets, which required targeted data augmentation. To address these imbalances, we employed GPT-4o to generate synthetic examples, particularly for low-resource language–class pairs. The generation prompts were designed with multilingual awareness, incorporating linguistic features, culturally appropriate idioms, and syntactic norms of each target language to enhance the realism and contextual alignment of the synthetic data. Augmentation outputs were rigorously filtered to ensure semantic validity, label fidelity, and language isolation, thereby preventing unintended language leakage or contamination. The resulting multilingual dataset was tokenized using a consistent scheme and validated for compatibility with the multilingual transformer models adopted for fine-tuning. This process enabled balanced exposure to both classes and promoted robust cross-lingual generalization.

3.2. Methodology

Our methodology aligns with the three evaluation settings defined by the shared task—monolingual, multilingual, and zero-shot generalization. We design our approach to explore both fine-tuning and prompting-based paradigms across high- and low-resource language scenarios.

Table 1
Distribution of Subjective (SUBJ) and Objective (OBJ) instances across monolingual datasets and the combined multilingual dataset. The monolingual rows represent filtered subsets of the training data specific to each language, while the multilingual row aggregates all available languages for joint training. This breakdown highlights the varying degrees of class imbalance and data availability across experimental settings

Setting	Language	SUBJ	OBJ	Total
Monolingual	Arabic	1055	1391	2446
	Bulgarian	312	379	691
	English	298	532	830
	German	308	492	800
	Italian	382	1231	1613
Multilingual	All Languages	2355	4025	6380

3.2.1. Fine-Tuning of Transformer Models (Settings 1 & 2)

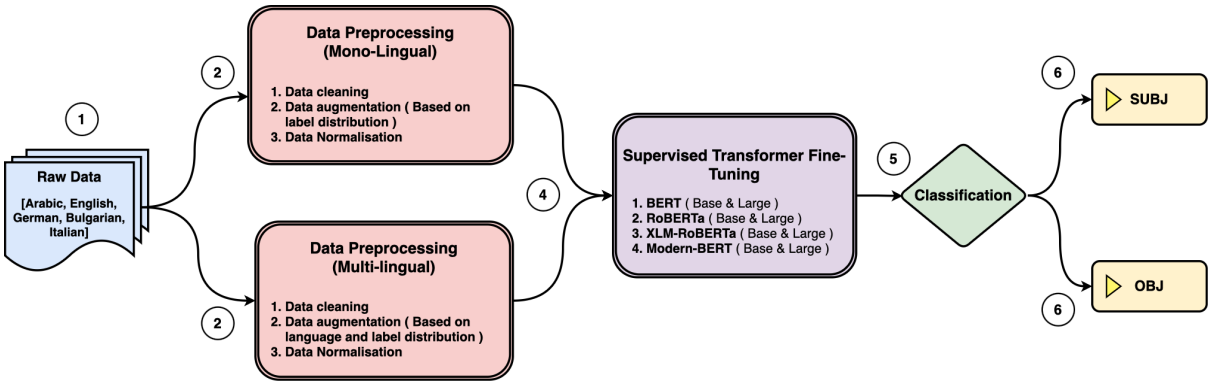


Figure 1: Architecture diagram illustrating the supervised transformer fine-tuning process for monolingual and multilingual setups.

For both the monolingual and multilingual settings, we employ supervised fine-tuning of pre-trained transformer-based language models. The following models were used in our experiments:

- **BERT**: BERT-Base and BERT-Large [49]
- **RoBERTa**: RoBERTa-Base and RoBERTa-Large [50]
- **XLM-RoBERTa**: XLM-RoBERTa-Base and XLM-RoBERTa-Large [54]
- **Modern-BERT**: Modern-BERT-Base and Modern-BERT-Large [67]

These models are known for their strong performance in cross-lingual and binary classification tasks. We fine-tune them using an augmented version of the training set and evaluate them on the dev-test split. The training loss is defined using the standard cross-entropy objective:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{1}$$

where $y_i \in \{0, 1\}$ is the true label (SUBJ or OBJ), and $\hat{y}_i$ is the predicted probability.
In the multilingual setting, the training set includes examples from all available languages, whereas in the monolingual setting, language-specific subsets are isolated for both training and evaluation.

3.2.2. In-Context Learning and Dynamic Prompting Framework (Setting 3)

To address zero-shot generalization for unseen languages (e.g., Ukrainian, Polish, Romanian, Greek), we employ In-Context Learning (ICL) using large-scale open-source language models. We begin with a

zero-shot inference setup using the Qwen-3-32B model, where the prompt consists solely of the query instance:

$$P = x\_q \tag{2}$$

This formulation relies entirely on the pretrained knowledge of the model to perform binary classification (SUBJ or OBJ), without providing any labeled support examples.
Initial results from zero-shot ICL showed moderate performance, but were limited by domain and language mismatch. To mitigate this, we extend the framework with a dynamic few-shot prompting strategy using a teacher–student architecture:

- **Student model** (`Qwen-2.5-3B`): Generates pseudo-labeled training data in the target (unseen) languages.
- **Teacher model** (`Qwen-3-32B`): Scores and filters the generated samples based on label consistency and semantic alignment.

The filtered samples form a high-quality candidate pool from which few-shot examples are selected dynamically for each test input. This selection is driven by cosine similarity between the sentence embeddings:

$$s(x\_q, x\_j) = \cos\left(\phi(x\_q), \phi(x\_j)\right) \tag{3}$$

where $\phi(x)$ denotes the embedding of input x. For every query x_q, the top-k most similar support examples x_j are retrieved to construct a contextual prompt.
This adaptive approach improves zero-shot generalization by incorporating semantically coherent examples, mitigating language shift, and simulating low-resource supervision through synthetic data curation.

4. Experimental Results

4.1. Setup

Model fine-tuning was conducted using high-performance GPUs to support efficient training across large multilingual datasets. Specifically, we utilized an NVIDIA RTX 4090 with 24 GB VRAM and an NVIDIA A100 with 40 GB VRAM. This hardware configuration enabled effective fine-tuning of transformer-based models with large parameter sizes and facilitated experimentation with various hyperparameter and resampling strategies under different task settings.

In addition to training infrastructure, we employed an NVIDIA H100 GPU for locally hosting large-scale models via the vLLM inference engine. This setup enabled rapid and cost-effective evaluation of models in real-time settings, including response generation, confidence scoring, and hybrid inference with ensemble strategies. The local hosting capability was crucial for integrating trained models into downstream verification pipelines with minimal latency.

This heterogeneous compute environment ensured both training efficiency and deployment scalability across the various stages of our experimentation.

4.2. Results

Evaluation Settings

We evaluate the performance of all transformer-based models under three primary settings:

- **Monolingual Fine-Tuning:** Models are fine-tuned and evaluated on individual language datasets. This setting assesses language-specific performance in resource-constrained conditions.
- **Multilingual Fine-Tuning:** Models are fine-tuned on an aggregated dataset comprising multiple languages (English, Arabic, Bulgarian, German, and Italian). This setup evaluates cross-lingual learning and robustness in a unified multilingual framework.
- **Zero-Shot Transfer:** Models fine-tuned in the multilingual setting are directly evaluated on unseen languages (Polish, Ukrainian, Greek, and Romanian) without any additional training. This setting examines the models' generalization capabilities to languages not seen during fine-tuning.

All evaluations are performed on the official dev-test splits provided as part of the shared task.

Monolingual Fine-Tuning

Table 2 presents the F1 scores obtained when models are trained and evaluated on individual languages. This setting reflects each model's capability to learn subjectivity and objectivity distinctions in a language-specific, resource-constrained environment.

Table 2
F1 scores for monolingual fine-tuning on individual languages. Evaluation is performed on the dev-test split for each respective language.

Model Name	English	Arabic	Bulgarian	German	Italian
BERT-Base	0.4852	0.5521	0.3639	0.5734	0.4196
BERT-Large	0.6406	0.5757	0.6242	0.6065	0.6794
RoBERTa-Base	0.4279	0.3680	0.3639	0.4058	0.4739
RoBERTa-Large	0.6687	0.5708	0.4353	0.7135	0.7205
XLM-RoBERTa-Base	0.4279	0.3623	0.3639	0.5373	0.4323
XLM-RoBERTa-Large	0.4935	0.4783	0.7412	0.7298	0.4879
Modern-BERT-Base	0.4723	0.5640	0.3879	0.5879	0.4327
Modern-BERT-Large	0.4924	0.5599	0.4432	0.6973	0.4923

Multilingual Fine-Tuning

Table 3 reports the F1 scores achieved by each model when trained on the combined dataset containing all five languages. This setup is designed to assess cross-lingual learning capabilities and robustness in a unified multilingual framework.

Table 3
F1 scores for multilingual fine-tuning using the combined dataset (Arabic + English + Bulgarian + German + Italian). Evaluated on the shared task dev-test split.

Model Name	Combined F1 Score
BERT-Base	0.6008
BERT-Large	0.6344
RoBERTa-Base	0.6249
RoBERTa-Large	0.6437
XLM-RoBERTa-Base	0.6443
XLM-RoBERTa-Large	0.6753
Modern-BERT-Base	0.5698
Modern-BERT-Large	0.6278

Zero-Shot Transfer to Unseen Languages

Table 4 presents the F1 scores for subjectivity claim classification in unseen languages where the model was not fine-tuned directly. This setting evaluates the zero-shot transfer capabilities of the models when applied to languages not present in the training set.

Table 4
F1 scores for zero-shot transfer to unseen languages (Setting 3). These languages were not included in the training data.

Language	F1 Score
Polish	0.38
Ukrainian	0.47
Greek	0.33
Romanian	0.52

Discussion

We observe that large models generally outperform their base counterparts, indicating a consistent benefit from increased model capacity. Among all models, `XLM-RoBERTa-Large` achieves the highest multilingual F1 score (0.6753), while also performing robustly in individual language settings such as Bulgarian and German. Interestingly, monolingual fine-tuning leads to strong results in resource-rich settings (e.g., Italian), but exhibits performance drops in lower-resource languages. This motivates the use of cross-lingual and multilingual pretraining as a means to mitigate language imbalance and improve generalization. In the zero-shot transfer setting, performance across unseen languages such as Greek and Polish remains modest, reflecting the challenge of applying pretrained models directly to languages not encountered during training. Since no prompt optimization or language-specific adaptation was applied, these results serve as a baseline for evaluating zero-shot capability. An alternative approach could involve translating inputs from unseen languages into a high-resource language (e.g., English), followed by classification using a monolingually fine-tuned model. Although translation may introduce noise, it could offer improved performance over direct zero-shot inference. Future work may explore such translation-based strategies alongside prompt tuning or few-shot adaptation to better support underrepresented languages.

5. Conclusion

In this work, we presented a robust approach for distinguishing subjective from objective content as part of Task 1 in the CLEF 2025 CheckThat! Lab. By fine-tuning transformer-based models and addressing class imbalance through targeted resampling and weighting techniques, our system achieves consistent performance across monolingual, multilingual, and zero-shot evaluation settings. Subjectivity detection serves as a crucial preliminary step in automated fact-checking pipelines by helping to identify opinionated or biased statements that require further scrutiny. Our method's strong adaptability to low-resource and cross-lingual scenarios demonstrates the effectiveness of leveraging multilingual pretrained representations for this task. Detailed error analysis further highlighted linguistic and contextual nuances influencing classification outcomes. Overall, our findings underscore the importance of multilingual and balanced data-driven modeling in enhancing the reliability of fact-checking systems and combating the spread of digital misinformation.

6. Future Work

Future research could focus on integrating subjectivity detection with downstream fact-checking components such as claim extraction and evidence retrieval to develop more comprehensive verification

pipelines. Expanding the model's coverage to additional languages and dialects would increase its global applicability. A key direction is developing models that achieve improved generalization and robustness on unseen languages, enhancing zero-shot cross-lingual transfer capabilities. Exploring advanced methods to address class imbalance, including adaptive loss functions and data augmentation, may further improve performance on less frequent classes. Incorporating richer contextual and pragmatic features, such as discourse relations and source reliability, could improve detection of nuanced subjectivity. Additionally, adopting continual learning and domain adaptation strategies would help maintain effectiveness amid evolving misinformation trends and new content domains.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and ChatGPT-4 in order to: perform grammar and spelling check, improve writing style and paraphrase and reword sections for clarity and conciseness. After using these tool(s)/service(s), the author(s) thoroughly reviewed, critically evaluated and edited all content to ensure accuracy and alignment with research objectives. The author(s) take(s) full responsibility for the publication's content.

References

[1] J. M. Wiebe, Tracking point of view in narrative, Computational Linguistics 20 (1994) 233–287. URL: https://aclanthology.org/J94-2004/.

[2] B. Pang, L. Lee, 2008. doi:10.1561/1500000011.

[3] A. Banfield, Unspeakable Sentences (Routledge Revivals): Narration and Representation in the Language of Fiction, 1st ed., 1982.

[4] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (Eds.), Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 347–354. URL: https://aclanthology.org/H05-1044/.

[5] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation (formerly Computers and the Humanities) 39 (2005) 164–210. doi:10.1007/s10579-005-7880-9.

[6] B. Liu, Sentiment analysis and opinion mining, volume 5, 2012. doi:10.2200/S00416ED1V01Y201204HLT016.

[7] V. Hatzivassiloglou, J. M. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity, in: COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics, 2000. URL: https://aclanthology.org/C00-1044/.

[8] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 129–136. URL: https://aclanthology.org/W03-1017/.

[9] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, volume 3406, 2005, pp. 486–497. doi:10.1007/978-3-540-30586-6_53.

[10] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112. URL: https://aclanthology.org/W03-1014/.

[11] P. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 417–424. URL: https://aclanthology.org/P02-1053/. doi:10.3115/1073083.1073153.

[12] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, 2002, pp. 79–86. URL: https://aclanthology.org/W02-1011/. doi:10.3115/1118693.1118704.

[13] K. Dave, S. Lawrence, D. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews 775152 (2003). doi:10.1145/775152.775226.

[14] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2015. doi:10.1017/CBO9781139084789.

[15] S. M. Mohammad, 9 - sentiment analysis: Detecting valence, emotions, and other affectual states from text, in: H. L. Meiselman (Ed.), Emotion Measurement, Woodhead Publishing, 2016, pp. 201–237. URL: https://www.sciencedirect.com/science/article/pii/B9780081005088000096. doi:https://doi.org/10.1016/B978-0-08-100508-8.00009-6.

[16] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis : A survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2018). doi:10.1002/widm.1253.

[17] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, V. Subrahmanian, Sentiment analysis: Adjectives and adverbs are better than adjectives alone, 2007. 2007 International Conference on Weblogs and Social Media, ICWSM 2007 ; Conference date: 26-03-2007 Through 28-03-2007.

[18] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2011) 267–307. URL: https://aclanthology.org/J11-2001/. doi:10.1162/COLI_a_00049.

[19] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: A. Zaenen, A. van den Bosch (Eds.), Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 976–983. URL: https://aclanthology.org/P07-1123/.

[20] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: M. Lapata, H. T. Ng (Eds.), Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 127–135. URL: https://aclanthology.org/D08-1014/.

[21] E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12 (2009) 526–558. doi:10.1007/s10791-008-9070-z.

[22] A. Balahur, M. Turchi, Multilingual sentiment analysis using machine translation?, in: A. Balahur, A. Montoyo, P. M. Barco, E. Boldrini (Eds.), Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, Jeju, Korea, 2012, pp. 52–60. URL: https://aclanthology.org/W12-3709/.

[23] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, J. Belyaeva, Sentiment analysis in the news, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: https://aclanthology.org/L10-1623/.

[24] F. Hamborg, K. Donnay, B. Gipp, Automated identification of media bias in news articles: an interdisciplinary literature review, International Journal on Digital Libraries 20 (2019). doi:10.1007/s00799-018-0261-y.

[25] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky, Linguistic models for analyzing and detecting biased language, in: H. Schuetze, P. Fung, M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1650–1659. URL: https://aclanthology.org/P13-1162/.

[26] C. Hube, B. Fetahu, Detecting biased statements in wikipedia, 2018, pp. 1779–1786. doi:10.1145/3184558.3191640.

[27] E. Baumer, E. Elovic, Y. Qin, F. Polletta, G. Gay, Testing and comparing computational approaches for identifying the language of framing in political news, in: R. Mihalcea, J. Chai, A. Sarkar

(Eds.), Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1472–1482. URL: https://aclanthology.org/N15-1171/. doi:10.3115/v1/N15-1171.

[28] M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in: K. Toutanova, H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 1113–1122. URL: https://aclanthology.org/P14-1105/. doi:10.3115/v1/P14-1105.

[29] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: https://aclanthology.org/L10-1263/.

[30] E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: The good the bad and the omg!, 2011.

[31] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[32] P. N. Ahmad, A. Shah, K. Lee, Enhanced propaganda detection in public social media discussions using a fine-tuned deep learning model: A diffusion of innovation perspective, Future Internet 17 (2025) 212. doi:10.3390/fi17050212.

[33] M. Hu, B. Liu, Mining and summarizing customer reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004). URL: https://api.semanticscholar.org/CorpusID:207155218.

[34] X. Ding, B. Liu, P. Yu, A holistic lexicon-based approach to opinion mining, 2008, pp. 231–240. doi:10.1145/1341531.1341561.

[35] K. Eguchi, V. Lavrenko, Sentiment retrieval using generative models, in: D. Jurafsky, E. Gaussier (Eds.), Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 345–354. URL: https://aclanthology.org/W06-1641/.

[36] M. Zhang, X. Ye, A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 411–418. URL: https://doi.org/10.1145/1390334.1390405. doi:10.1145/1390334.1390405.

[37] L. Soldaini, E. Yom-Tov, Inferring individual attributes from search engine queries and auxiliary information, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 293–301. URL: https://doi.org/10.1145/3038912.3052629. doi:10.1145/3038912.3052629.

[38] R. White, E. Horvitz, Cyberchondria: Studies of the escalation of medical concerns in web search, ACM Trans. Inf. Syst. 27 (2009). doi:10.1145/1629096.1629101.

[39] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3346–3359. URL: https://aclanthology.org/C18-1283/.

[40] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7740–7754. URL: https://aclanthology.org/2020.emnlp-main.623/. doi:10.18653/v1/2020.emnlp-main.623.

[41] T. Alhindi, S. Petridis, S. Muresan, Where is your evidence: Improving fact-checking by justification

modeling, in: J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal (Eds.), Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 85–90. URL: https://aclanthology.org/W18-5513/. doi:10.18653/v1/W18-5513.

[42] R. F. Bruce, J. M. Wiebe, Recognizing subjectivity: a case study in manual tagging, Nat. Lang. Eng. 5 (1999) 187–205. URL: https://doi.org/10.1017/S1351324999002181. doi:10.1017/S1351324999002181.

[43] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, Learning subjective language, Comput. Linguist. 30 (2004) 277–308. URL: https://doi.org/10.1162/0891201041850885. doi:10.1162/0891201041850885.

[44] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 271–278. URL: https://aclanthology.org/P04-1035/. doi:10.3115/1218955.1218990.

[45] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (2008) 1–135. URL: https://doi.org/10.1561/1500000011. doi:10.1561/1500000011.

[46] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, in: C. Callison-Burch, S. Wan (Eds.), Proceedings of the ACL Student Research Workshop, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 43–48. URL: https://aclanthology.org/P05-2008/.

[47] L. Zhang, B. Liu, Identifying noun product features that imply opinions, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 575–580. URL: https://aclanthology.org/P11-2101/.

[48] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181/. doi:10.3115/v1/D14-1181.

[49] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[50] 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, 2019. doi:10.48550/arXiv.1907.11692.

[51] T. Mikolov, G. Corrado, K. Chen, J. Dean, Efficient estimation of word representations in vector space, 2013, pp. 1–12.

[52] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: L. Vanderwende, H. Daumé III, K. Kirchhoff (Eds.), Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751. URL: https://aclanthology.org/N13-1090/.

[53] A. Conneau, G. Lample, Cross-lingual language model pretraining, Curran Associates Inc., Red Hook, NY, USA, 2019.

[54] 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.

[55] P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. Martino, P. Nakov, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, 2018. doi:10.48550/arXiv.1808.05542.

[56] C. Hansen, C. Hansen, J. G. Simonsen, C. Lioma, Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_56.pdf.

[57] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. S. Ali, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2020, p. 215–236. URL: https://doi.org/10.1007/978-3-030-58219-7_17. doi:10.1007/978-3-030-58219-7_17.

[58] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, andnbsp;fake news, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 264–291. URL: https://doi.org/10.1007/978-3-030-85251-1_19. doi:10.1007/978-3-030-85251-1_19.

[59] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2022, p. 416–428. URL: https://doi.org/10.1007/978-3-030-99739-7_52. doi:10.1007/978-3-030-99739-7_52.

[60] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, andnbsp;authority, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, p. 506–517. URL: https://doi.org/10.1007/978-3-031-28241-6_59. doi:10.1007/978-3-031-28241-6_59.

[61] R. Frick, I. Vogel, Fraunhofer sit at checkthat! 2023: Mixing single-modal classifiers to estimate the check-worthiness of multi-modal tweets, 2023. doi:10.48550/arXiv.2307.00610.

[62] 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/.

[63] 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: D. Byron, A. Venkataraman, D. Zhang (Eds.), Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 34–35. URL: https://aclanthology.org/H05-2018/.

[64] O. Tsur, A. Rappoport, What's in a hashtag? content based prediction of the spread of ideas in microblogging communities, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 643–652. URL: https://doi.org/10.1145/2124295.2124320. doi:10.1145/2124295.

2124320.

[65] B. Plank, A. Søgaard, Y. Goldberg, Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 412–418. URL: https://aclanthology.org/P16-2067/. doi:10.18653/v1/P16-2067.

[66] T. Hercig, P. Kral, Evaluation datasets for cross-lingual semantic textual similarity, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 524–529. URL: https://aclanthology.org/2021.ranlp-1.59/.

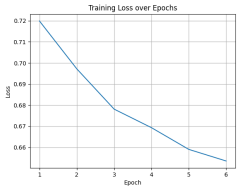
[67] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

A. Training Loss Plots

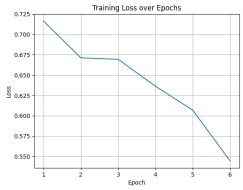
In this appendix, we present training loss plots for each language–model combination.

A.1. Mono-Lingual Loss Plots

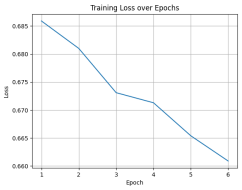
A.1.1. Arabic



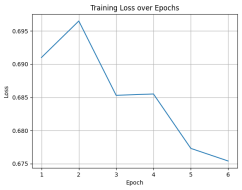
(a) ModernBERT-base



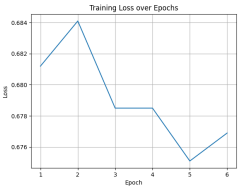
(b) ModernBERT-large



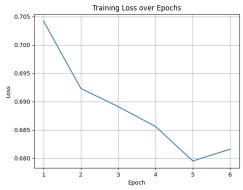
(c) BERT-base



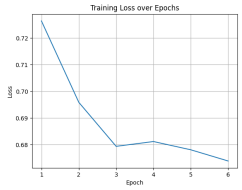
(d) BERT-large



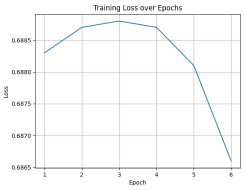
(e) RoBERTa-base



(f) RoBERTa-large



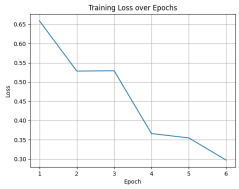
(g) XLM-RoBERTa-base



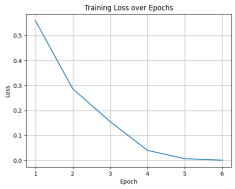
(h) XLM-RoBERTa-large

Figure 2: Training loss plots for Arabic models.

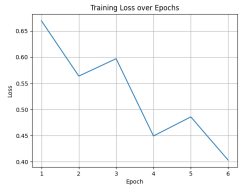
A.1.2. English



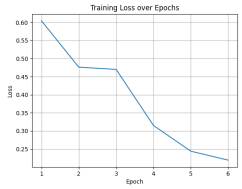
(a) ModernBERT-base



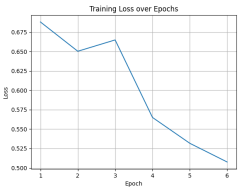
(b) ModernBERT-large



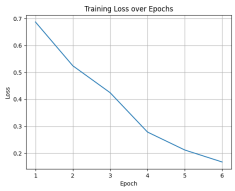
(c) BERT-base



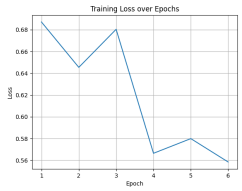
(d) BERT-large



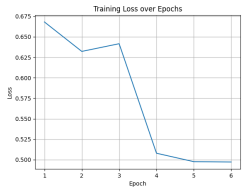
(e) RoBERTa-base



(f) RoBERTa-large



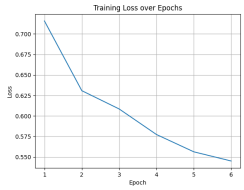
(g) XLM-RoBERTa-base



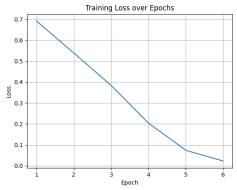
(h) XLM-RoBERTa-large

Figure 3: Training loss plots for English models.

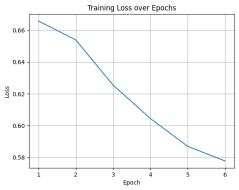
A.1.3. German



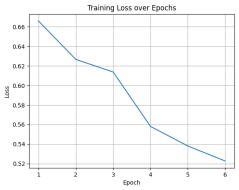
(a) ModernBERT-base



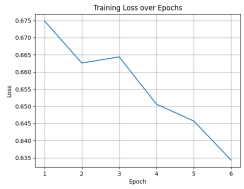
(b) ModernBERT-large



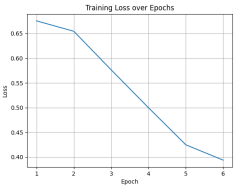
(c) BERT-base



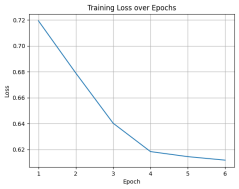
(d) BERT-large



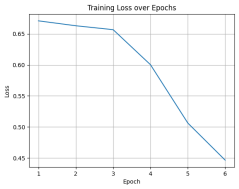
(e) RoBERTa-base



(f) RoBERTa-large



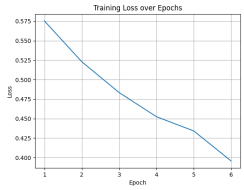
(g) XLM-RoBERTa-base



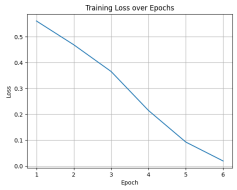
(h) XLM-RoBERTa-large

Figure 4: Training loss plots for German models.

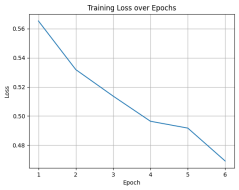
A.1.4. Italian



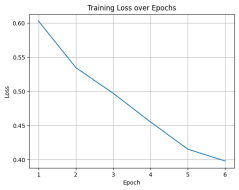
(a) ModernBERT-base



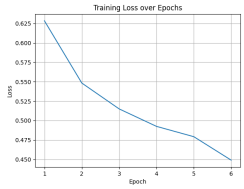
(b) ModernBERT-large



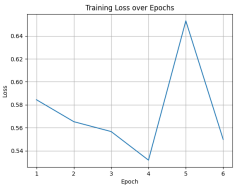
(c) BERT-base



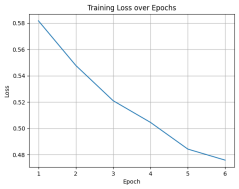
(d) BERT-large



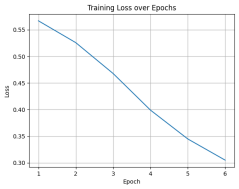
(e) RoBERTa-base



(f) RoBERTa-large



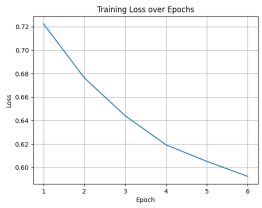
(g) XLM-RoBERTa-base



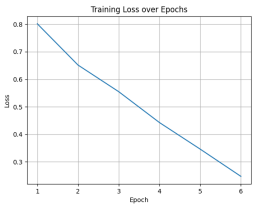
(h) XLM-RoBERTa-large

Figure 5: Training loss plots for Italian models.

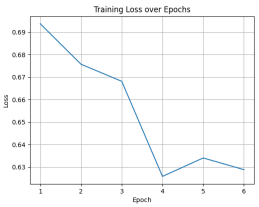
A.1.5. Bulgarian



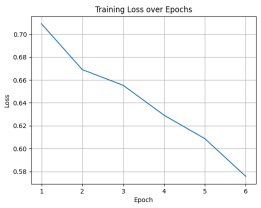
(a) ModernBERT-base



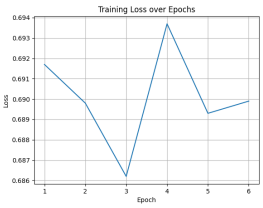
(b) ModernBERT-large



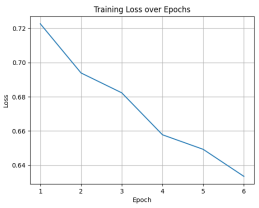
(c) BERT-base



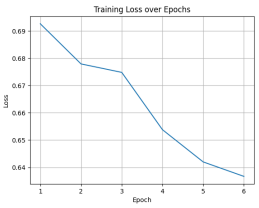
(d) BERT-large



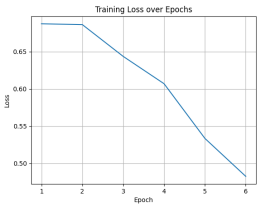
(e) RoBERTa-base



(f) RoBERTa-large



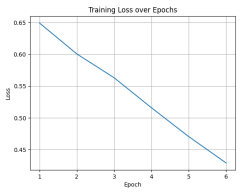
(g) XLM-RoBERTa-base



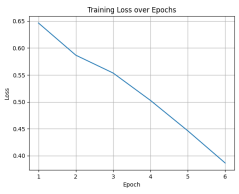
(h) XLM-RoBERTa-large

Figure 6: Training loss plots for Bulgarian models.

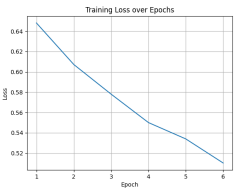
A.2. Multi-Lingual Loss Plots



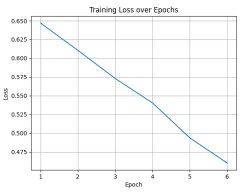
(a) ModernBERT-base



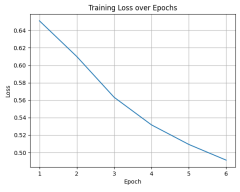
(b) ModernBERT-large



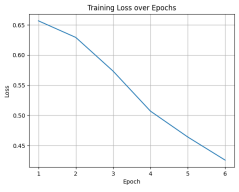
(c) BERT-base



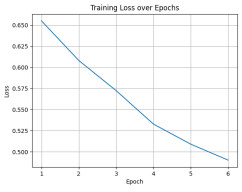
(d) BERT-large



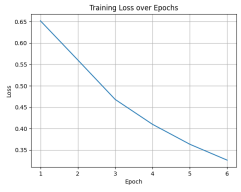
(e) RoBERTa-base



(f) RoBERTa-large



(g) XLM-RoBERTa-base



(h) XLM-RoBERTa-large

Figure 7: Training loss plots for combined multilingual models.