Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

RoIt-XMASA: Multi-Domain Multilingual Sentiment Analysis Dataset for Romanian and Italian

Andrei-Marius Avram, Aureliu-Valentin Antonie*, Cosmin-Mircea Croitoru*, Vlad-Andrei Muntean*, Dumitru-Clementin Cercel†

National University of Science and Technology POLITEHNICA Bucharest,
Bucharest, Romania
dumitru.cercel@upb.ro

Abstract

We present RoIt-XMASA, a multilingual dataset that extends the Cross-lingual Multi-domain Amazon Sentiment Analysis to Italian and Romanian, comprising 36,000 labeled reviews across three domains (books, movies, and music) and 202,141 unlabeled samples. To address cross-lingual and cross-domain challenges, we propose a multi-target adversarial training framework that employs loss reversal with meta-learned coefficients to dynamically balance sentiment discrimination with domain and language invariance. XLM-R achieves an F1-score of 66.23% with our approach, outperforming the baseline by 4.64%. Few-shot evaluation shows that Llama-3.1-8B achieves 58.43% F1-score, revealing a meaningful trade-off between the efficiency of prompting-based approaches and the higher performance of task-specific fine-tuning.

Introduction

Cross-domain and cross-lingual sentiment analysis remains a fundamental challenge, requiring models to generalize across both linguistic and topical boundaries. Although multilingual models have advanced significantly, the double shift in language and domain, compounded by the scarcity of resources for languages such as Romanian, presents a major obstacle (Zhao et al. 2024). These limitations emphasize the need for frameworks that leverage large-scale unlabeled data and adversarial learning (Ganin et al. 2016; Chen et al. 2018) to extract invariant sentiment features across disparate distributions.

We introduce RoIt-XMASA, a large-scale multilingual dataset extending the Cross-lingual Multi-domain Amazon Sentiment Analysis (XMASA) corpus (Blitzer, Dredze, and Pereira 2007) to Italian and Romanian. Our dataset comprises 36,000 annotated reviews equally distributed across three domains (books, movies, and music) and two languages, with an additional 202,141 unlabeled samples for semi-supervised learning. Each review is annotated with ratings from 1 to 5 stars (excluding 3), maintaining the original XMASA structure while adapting to the linguistic characteristics of Italian and Romanian.

Our methodological contribution introduces a multi-target adversarial training framework (Ganin et al. 2016; Chen et al. 2018) that simultaneously optimizes for sentiment classification while learning representations invariant to domain and language attributes. We employ loss reversal (Avram et al. 2024) rather than gradient reversal and dynamically adjust adversarial coefficients through meta-learning (Vettoruzzo et al. 2024), thus eliminating the need for manual hyperparameter tuning.

Experiments with multilingual variants of BERT (Devlin et al. 2019) demonstrate that our approach obtains substantial improvements: XLM-R (Conneau et al. 2020) achieves an F1-score of 66.23% (an improvement of 4.64% over the baseline) when using both domain and language as adversarial targets. We also establish baselines using recent open-source large language models (LLMs), with Llama-3.1-8B (Dubey et al. 2024) reaching a 58.43% F1-score in few-shot settings.

Furthermore, we investigate the utility of the 202,141 unlabeled samples in RoIt-XMASA for unsupervised domain adaptation. By performing a computationally efficient fine-tuning phase with Low-Rank Adaptation (LoRA) (Hu et al. 2022) for a single epoch, we assess how brief exposure to domain-specific unlabeled data in Romanian and Italian can bridge the gap between general-purpose pre-training and the specific linguistic registers of e-commerce reviews. This adaptive step aims to improve the model alignment with the target distributions before the final classification task.

The main contributions of our work can be summarized as follows:

- We present RoIt-XMASA, the first large-scale extension of the XMASA dataset (Blitzer, Dredze, and Pereira 2007) to Italian and Romanian, providing 36,000 labeled and 202,141 unlabeled reviews balanced across domains and languages[1].
- We propose a multi-target adversarial training framework with meta-learned coefficients that achieves consistent improvements across all tested multilingual models.
- We establish comprehensive baselines using both fine-tuned encoder models and few-shot LLM approaches,

*Equal contribution. Alphabetical author order.
†Corresponding author.

[1]The dataset is available at the following link: https://huggingface.co/datasets/avramandrei/RoIt-XMASA

revealing distinct linguistic patterns between Italian and Romanian reviewing styles with implications for cross-cultural natural language processing (NLP) applications.

- We validate the utility of the unlabeled subset by demonstrating that an unsupervised adaptation via LoRA measurably improves LLM performance, effectively aligning models with target languages and domains.

Related Work

Evolution of Multilingual Sentiment Analysis

Multilingual sentiment analysis has evolved significantly from early methods that relied on machine translation (Araujo et al. 2016) and lexicon-based features (Qi and Shabrina 2023). Initial cross-lingual approaches often involved translating test data into a high-resource language such as English (a technique known as a translate-test) or training classifiers on translated source data (Prettenhofer and Stein 2010; Wan 2009). Another line of work focused on creating multilingual sentiment-aware word embeddings, either by aligning monolingual vector spaces or learning them jointly from parallel corpora, sometimes using pivot languages to bridge resource gaps (Xu and Wan 2017). We note that these methods frequently suffer from translation errors and the loss of cultural nuances.

The advent of pre-trained language models marked a paradigm shift in the field. Multilingual models such as mBERT (Devlin et al. 2019) and XLM-R (Conneau et al. 2020) have demonstrated a remarkable capacity for zero-shot cross-lingual transfer by learning shared representations in more than 100 languages. Subsequent research has extensively benchmarked these models, confirming their effectiveness and also highlighting performance disparities across languages and tasks (Rajda et al. 2022; Augustyniak et al. 2023). Models are typically evaluated as static feature extractors for a downstream classifier or via full fine-tuning, with the latter generally obtaining superior performance but at a higher computational cost.

Despite the success of large-scale multilingual models, significant challenges remain, particularly for low-resource languages and culture-dependent tasks. For example, sentiment expression is deeply intertwined with cultural context, and models pre-trained on generic web corpora may fail to capture fine-grained subtleties (Augustyniak et al. 2023). This has spurred efforts to create specialized and high-quality datasets for underrepresented languages, such as the NusaX corpus for Indonesian dialects (Winata et al. 2023). Such resources are critical for developing models that are not only linguistically competent but also culturally aware, motivating the creation of our dataset, RoIt-XMASA.

Cross-Lingual Adaptation and Robustness

To improve the model robustness across languages and domains, researchers have explored adversarial training techniques (Chen et al. 2018). The core idea is to learn feature representations that are discriminative for the primary task (e.g., sentiment classification) but invariant to nuisance variables like the language or domain of the input text. This is often achieved by introducing a domain classifier that is trained to predict the nuisance variable from the learned features, while the main feature extractor is trained to fool this classifier, typically through a gradient reversal layer (Ganin et al. 2016; Ye et al. 2020). This forces the model to learn more generalized and transferable features, making it a powerful technique for cross-lingual and cross-domain adaptation.

Most recently, the focus has shifted toward LLMs and their ability to perform sentiment tasks via prompting or instruction tuning. Although models like Llama and Qwen exhibit strong zero-shot capabilities, their performance in specialized domains or low-resource languages often benefits from targeted adaptation (Hu et al. 2022). Parameter-efficient techniques, particularly LoRA, have emerged as a standard for adapting these massive models without the prohibitive costs of full-parameter updates. Recent studies suggest that performing an initial phase of unsupervised domain adaptation—where the model is exposed to unlabeled target-domain text—can significantly improve alignment and subsequent classification accuracy (Saad-Falcon et al. 2023; Zhang et al. 2024; Xing et al. 2025). Our work builds on this by quantifying the impact of such unlabeled adaptation on the Romanian and Italian review domains, positioning it as a complementary strategy to adversarial training in multi-domain scenarios.

RoIt-XMASA Dataset

RoIt-XMASA, the multilingual dataset we introduce in this work, extends the XMASA dataset (Blitzer, Dredze, and Pereira 2007) to Italian and Romanian languages, maintaining the three-domain structure (books, movies, and music) while adapting to the linguistic and cultural characteristics of these languages. Our dataset contains 36,000 annotated reviews crawled from the Internet, which were equally distributed across all dimensions.

Dataset Collection

The construction of the corpus followed a language-specific acquisition strategy to ensure high-quality and authentic data. The Italian reviews were collected from Amazon to maintain consistency with the original source of XMASA. In contrast, due to the lack of Romanian-language reviews on Amazon, the Romanian reviews were acquired from a combination of representative local platforms and manually translated English sources to better capture native linguistic and cultural patterns for each of the three domains: (1) book reviews were collected from Goodreads[2] and Audiotribe[3], with additional samples manually translated from XMASA's English data to reach the target domain size of 6k reviews, (2) movie reviews were gathered from Cinemagia[4], a prominent Romanian film review platform, and (3) music reviews were obtained from XMASA's English data and manually translated into Romanian.

The dataset follows a balanced design with 6,000 labeled samples in each of the train, validation, and test splits. Table

[2]https://www.goodreads.com/

[3]https://audiotribe.ro

[4]https://www.cinemagia.ro/

Sets	Books		Movies		Music	
	IT	RO	IT	RO	IT	RO
Train	2k	2k	2k	2k	2k	2k
Valid	2k	2k	2k	2k	2k	2k
Test	2k	2k	2k	2k	2k	2k
Total labeled	**6k**	**6k**	**6k**	**6k**	**6k**	**6k**

Table 1: It-XMASA and Ro-XMASA dataset statistics by domain and language on the train, validation, and test subsets.

1 presents the distribution across domains and languages. Each language-domain combination contains exactly 2,000 samples per split, ensuring balanced representation for both cross-lingual and cross-domain evaluation scenarios.

In addition to the labeled splits, we provide an unlabeled dataset of 202,141 samples to facilitate research on semi-supervised learning and domain adaptation. The unlabeled data set spans all three domains: Italian samples include 14,722 books, 32,970 movies, and 53,662 music reviews, while Romanian samples comprise 13,429 books, 60,023 movies, and 27,335 music reviews.

Rating Distribution

Figure 1 illustrates the rating distribution across the entire RoIt-XMASA dataset and its breakdown by language and domain. The dataset exhibits a naturalistic distribution with peaks at the extreme ratings (1 and 5 stars), reflecting the tendency of users to review products they either strongly like or dislike.

The distribution shows 10,341 one-star ratings, 7,639 two-star ratings, 6,540 four-star ratings, and 11,480 five-star ratings. This polarization pattern is consistent across both languages, with Italian showing a slightly more balanced distribution of intermediate ratings than Romanian. This pattern also appears in both the movies and the music domains.

Textual Characteristics

The RoIt-XMASA dataset exhibits some variation in the review length and linguistic properties. Overall, reviews contain a mean of 63.18 tokens and a median of 33, with a rather high variance (i.e., a standard deviation of 100.78), ranging from single-token reviews to extensive critiques of up to 3,391 tokens. This wide distribution, illustrated in Figure 2, reveals the diverse nature of user expressions in sentiment communication, with a characteristic long-tailed distribution typical of user-generated content.

The titles show even greater variability, with a mean of 2.20 tokens and a median of approximately 1 token, but a highly skewed distribution. The presence or absence of titles appears to be strongly influenced by cultural and platform-specific conventions, as revealed by differences across languages (i.e., out of the 36k collected samples, only 19,762 have titles). A comprehensive cross-language and cross-domain analysis of these textual characteristics is provided in the Appendix.

Data Quality

Quality assessment reveals that there are no duplicate entries in RoIt-XMASA when considering the combination of title and text fields. This ensures that each sample represents a unique review instance, preventing data leakage between splits and maintaining the integrity of experimental evaluations. The RoIt-XMASA dataset underwent rigorous validation and cleaning procedures to ensure data quality:

- **Language verification**: All reviews were validated using the `cld3`[5] library to ensure correct language assignment. Reviews with confidence scores below 0.95 or that were detected as different languages were manually reviewed, corrected, or removed.
- **HTML and special character removal**: Systematic cleaning removed all HTML tags, entities (e.g., &,), and hidden control characters that could interfere with model processing.
- **Text normalization**: Applied comprehensive normalization rules to standardize text representation while preserving semantic content. The complete normalization pipeline is detailed in the Appendix.

To validate the automated processing pipeline, we conducted manual verification by human annotators. Three native speakers for each language reviewed 100 randomly sampled reviews (200 total), assessing language correctness, sentiment-rating alignment, and text quality. The inter-annotator agreement achieved a Krippendorff's alpha of 0.82, indicating strong agreement and confirming the reliability of our data quality measures.

The balanced structure of the RoIt-XMASA dataset, the natural rating distribution, and various textual characteristics make it well-suited for investigating the sentiment analysis challenges in both the cross-lingual and cross-domain settings, extending the foundational work of Blitzer, Dredze, and Pereira (2007) to new linguistic territories while maintaining methodological consistency.

Methodology

Multi-Adversarial Objective with Meta-Learned Coefficients

We build on the multi-target adversarial training framework introduced by Avram et al. (2025a) and adapt it for sentiment classification. Let f_θ denote the shared encoder, and h_ϕ the task-specific classification head with task loss $\mathcal{L}_{\text{task}}$ for each of the three possible prediction objectives (i.e., rating, language, and domain). Then, in our setup, the rating prediction objective is designated as the primary task to be optimized, while the remaining two are treated as adversarial objectives.

Unlike previous work (Ganin et al. 2016; Chen et al. 2018) that employs gradient reversal, we adopt loss reversal, which directly flips the sign of adversarial objectives during optimization, shown in previous work to yield better results (Avram et al. 2024, 2025b). Thus, the overall loss is as follows:

[5]https://github.com/google/cld3

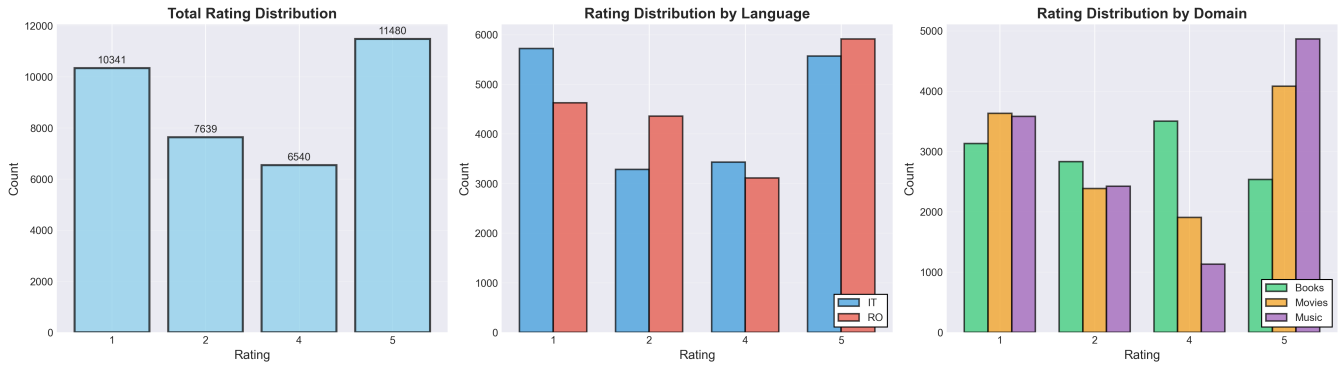

Figure 1: Rating distribution across the RoIt-XMASA dataset. Left: overall rating distribution; Center: rating distribution by language; Right: rating distribution by domain.

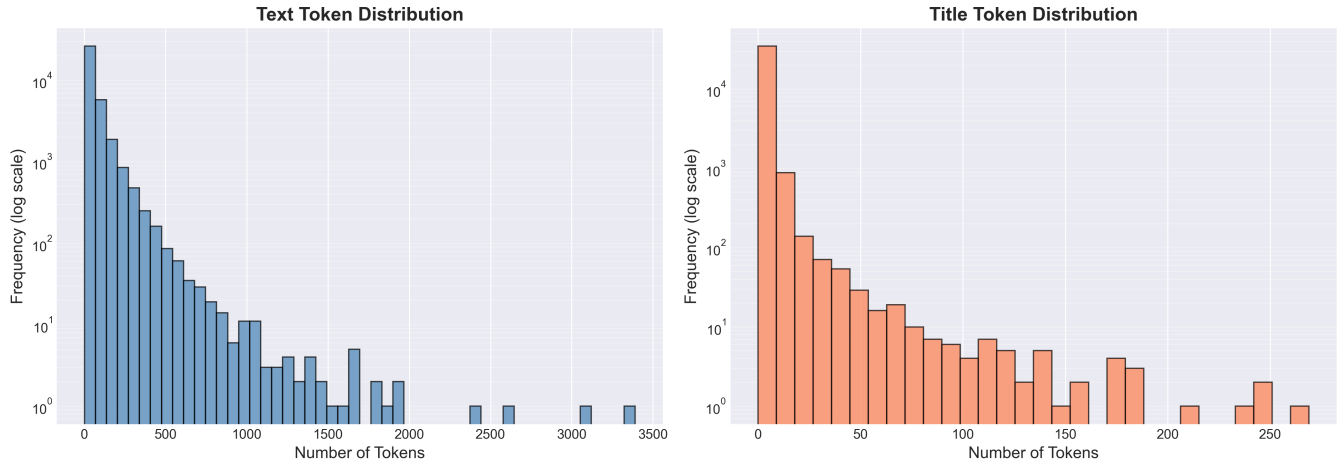

Figure 2: Token distribution histograms for the text and title fields.

$$\mathcal{L} = \mathcal{L}_{\text{rating}} - \lambda_1 \mathcal{L}_{\text{domain}} - \lambda_2 \mathcal{L}_{\text{lang}}, \quad (1)$$

where λ_1 and λ_2 are adversarial coefficients. To avoid manual tuning, we adopt a meta-learning strategy: coefficients are updated to minimize a validation-set meta-loss:

$$\lambda \leftarrow \lambda - \eta \nabla_\lambda \mathcal{L}_{\text{meta}}(\lambda), \quad (2)$$

with η denoting the meta-learning rate. This procedure dynamically balances the invariance to adversarial attributes with the discriminability of the primary task.

All multilingual encoder models, namely M-BERT (Devlin et al. 2019), XLM (Conneau and Lample 2019), and XLM-R (Conneau et al. 2020), were fine-tuned using the same training protocol to ensure a fair comparison across architectures. This includes identical optimization settings, batch sizes, learning rate schedules, and early stopping criteria. We report the complete set of hyperparameters used to fine-tune multilingual encoder models in the Appendix.

Open-Source LLM Baselines

We evaluate recent open-source large language models as baselines for cross-domain and cross-lingual sentiment classification. The selection criterion was to include models between 6B and 9B parameters, always using the latest version released for each family to ensure a representative and unbiased comparison. Concretely, we benchmark the few-shot performance of four open-source LLMs: Llama-3.1-8B[6], Qwen3-8B[7], Mistral-7B-v0.3[8], and DeepSeek-R1-8B[9].

Furthermore, to better align open-source models with the specific linguistic and domain distributions of our corpus, we performed an initial stage of fine-tuning on the unlabeled data. Using the 202,141 unlabeled samples available in RoIt-XMASA, we fine-tuned each LLM for one epoch using LoRA. This adaptation phase employed a causal language modeling objective, allowing the models to learn the nuances of Italian and Romanian review styles across the three domains without the computational overhead of full-parameter tuning.

All models were evaluated in a prompting-based setup,

[6]https://huggingface.co/meta-llama/Llama-3.1-8B
[7]https://huggingface.co/Qwen/Qwen3-8B
[8]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[9]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

Model	Dom.	Lang.	Books		Movies		Music		IT		RO		Avg.	
			Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
M-BERT	✗	✗	62.15	57.23	65.82	60.45	63.05	57.64	66.92	62.18	60.42	54.70	63.67	58.44
M-BERT	✓	✓	65.23	60.89	68.45	64.28	**67.91**	63.15	69.45	65.12	**65.28**	61.35	66.45	62.17
XLM	✗	✗	60.34	57.12	63.78	60.89	60.89	57.79	64.23	61.45	59.11	55.75	61.67	58.60
XLM	✓	✓	**66.78**	62.41	65.12	62.18	61.78	59.55	66.34	62.67	62.78	**61.42**	64.56	61.38
XLM-R	✗	✗	63.18	60.34	66.82	63.78	63.71	60.65	67.45	64.23	61.69	58.95	64.57	61.59
XLM-R	✓	✓	66.45	**64.12**	**71.12**	**68.61**	67.15	**65.78**	**72.42**	**69.40**	63.39	60.93	**68.91**	**66.23**

Table 2: Multi-adversarial training results on RoIt-XMASA. Down arrow (↓) indicates adversarial optimization. Domain and language columns show performance by domain (Books, Movies, Music) and language (IT=Italian, RO=Romanian).

with prompts and LLM experiment hyperparameters detailed in the Appendix.

Results

We present our experimental findings organized into three subsections: multi-adversarial training with multilingual encoders, few-shot performance of open-source LLMs, and the impact of unlabeled adaptive fine-tuning. All tables report performance disaggregated by domain and language, with the final column showing macro-averaged results.

Multi-Adversarial Training Performance

In this set of experiments, we apply our framework to three established multilingual models: M-BERT, XLM, and XLM-R. Table 2 shows that our multi-target adversarial training framework consistently improves sentiment classification performance across all baseline models. The most significant gains are observed when both domain and language are used as adversarial objectives, confirming that learning representations invariant to both factors is beneficial.

XLM-R emerges as the top-performing model overall, achieving an average F1-score of 66.23% with multi-adversarial training, a 4.64% improvement over its baseline of 61.59%. However, the model performance varies across domains and languages, and no single model dominates all settings. XLM-R excels at movies (68.61% F1-score) and Italian reviews (69.40% F1-score), while XLM-R leads both in book performance (64.12% F1-score) and in music performance (65.78% F1-score). Notably, M-BERT achieves the highest Romanian F1-score (61.35%), outperforming both XLM and XLM-R on this more challenging language. These complementary strengths suggest that ensemble approaches or language-specific model selection could further improve cross-lingual sentiment analysis.

LLM Baseline Performance

The evaluation results, summarized in Table 3, reveal the current capabilities of these models in a few-shot learning context for this task. Overall, performance improves with an increased number of in-context examples, although the gains are not always monotonic.

Llama-3.1-8B achieves the highest overall performance with a 58.43% F1-score in the 5-shot setting, excelling particularly on movies (61.85% F1-score). However, the model strengths vary by domain and language: DeepSeek-R1-8B demonstrates better performance on books (57.81% F1-score) and Romanian reviews (57.04% F1-score), while Qwen3-8B leads in music classification (59.34% F1-score). The cross-lingual gap persists across all models, though DeepSeek-R1-8B shows the smallest Italian-Romanian disparity (61.23% vs. 57.04% F1-scores), suggesting better multilingual calibration.

These results highlight an important trade-off between the two paradigms. Although the best fine-tuned encoder (XLM-R with multi-adversarial training at 66.23% F1-score) achieves notably higher performance than the best few-shot LLM (Llama-3.1-8B at 58.43% F1-score), the LLM-based approach requires no task-specific labeled data or gradient-based optimization, relying solely on a handful of in-context examples. This observation makes a few-shot LLM evaluation particularly attractive in low-resource scenarios where the labeled dataset is scarce or annotation is costly. Conversely, when sufficient labeled data and computational resources are available, task-specific fine-tuning with adversarial training remains the more performant choice, underscoring the complementary nature of both approaches given practical constraints.

Unlabeled Adaptive Fine-Tuning

To evaluate the impact of domain-specific adaptation without labeled task data, we conducted a series of experiments using unlabeled adaptive fine-tuning. In this setup, the base LLMs were fine-tuned for a single epoch on the 202,141 unlabeled samples of the RoIt-XMASA corpus using LoRA. This process focuses exclusively on language modeling the target domains (books, movies, and music) in Romanian and Italian, without any exposure to sentiment labels.

The results in Table 4 show consistent improvements across all models, though adaptation benefits vary by model and domain. Llama-3.1-8B gains 1.72 percentage points overall (58.43% $\rightarrow$ 60.15% F1), with the largest improvements in movies (+1.93 F1) and music (+2.30 F1). Qwen3-8B shows the strongest gains for Romanian reviews (+5.67 F1), suggesting particularly effective cross-lingual adaptation from the unlabeled corpus. DeepSeek-R1-8B exhibits strong book adaptation (+1.53 F1) and achieves the best Italian performance post-adaptation (62.89% F1), while Mistral-7B-v0.3 shows modest but consistent gains across

Model	# Shots	Books		Movies		Music		IT		RO		Avg.	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Llama-3.1-8B	0	58.23	49.12	62.45	53.68	60.08	50.08	63.12	54.23	57.34	47.69	60.25	50.96
Llama-3.1-8B	5	61.89	56.78	**66.23**	**61.85**	63.12	58.15	**66.34**	60.89	61.15	56.63	**63.66**	**58.43**
Qwen3-8B	0	56.45	48.23	60.34	52.78	58.05	49.26	61.12	52.89	55.44	47.29	58.28	50.09
Qwen3-8B	3	59.67	54.23	61.45	56.12	**63.78**	**59.34**	63.45	57.92	59.82	55.11	60.11	54.97
Mistral-7B-v0.3	0	52.34	46.12	56.23	50.12	53.97	47.85	56.78	50.23	51.58	45.83	54.18	48.03
Mistral-7B-v0.3	5	53.89	47.23	57.45	51.34	54.56	48.40	57.89	51.12	53.71	46.86	55.30	48.99
DeepSeek-R1-8B	0	59.78	54.89	63.67	58.67	61.29	55.49	64.45	58.92	58.71	53.78	61.58	56.35
DeepSeek-R1-8B	1	**62.45**	**57.81**	63.34	59.12	61.23	55.67	65.12	**61.23**	**61.89**	**57.04**	62.01	56.80

Table 3: Few-shot sentiment classification performance of open-source LLMs on RoIt-XMASA, showing performance by domain and language. Only baseline (0-shot) and best-performing few-shorts configurations are shown for each model.

Model	Config.	Books		Movies		Music		IT		RO		Avg.	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Llama-3.1-8B	Base	61.89	56.78	66.23	61.85	63.12	58.15	66.34	60.89	61.15	56.63	63.66	58.43
	Adapted	63.45	58.92	**68.12**	**63.78**	**65.28**	**60.45**	**67.89**	62.34	63.34	59.43	**65.22**	**60.15**
Qwen3-8B	Base	59.67	54.23	61.45	56.12	63.78	59.34	63.45	57.92	59.82	55.11	60.11	54.97
	Adapted	61.78	56.45	63.89	59.18	64.12	59.87	65.12	59.34	**64.56**	**60.78**	62.14	57.20
Mistral-7B-v0.3	Base	53.89	47.23	57.45	51.34	54.56	48.40	57.89	51.12	53.71	46.86	55.30	48.99
	Adapted	**56.23**	49.78	58.12	51.89	55.21	49.45	58.34	51.67	55.40	48.78	56.85	50.12
DeepSeek-R1-8B	Base	62.45	57.81	63.34	59.12	61.23	55.67	65.12	61.23	61.89	57.04	62.01	56.80
	Adapted	63.78	**59.34**	64.23	59.78	61.89	56.23	66.12	**62.89**	63.67	58.12	62.35	56.95

Table 4: Impact of 1-epoch LoRA adaptation on LLMs, showing base vs. adapted performance by domain and language. Results are reported using the best-performing few-shot configuration for each model.

all domains. These varied adaptation patterns demonstrate that brief unsupervised exposure to target domains and languages measurably enhances LLM performance, with model-specific strengths emerging through domain alignment.

Conclusion

This paper introduced RoIt-XMASA, an extension of the XMASA framework to Italian and Romanian, using 36,000 labeled reviews across three domains. Our multi-target adversarial training framework with meta-learned coefficients achieved significant improvements, with XLM-R reaching 66.23% F1-score when using both domain and language as adversarial targets. Few-shot LLM baselines, while lower in absolute performance, achieved competitive results without task-specific fine-tuning, revealing a practical trade-off between, on the one hand, labeling and training costs, and, on the other, classification accuracy. The dataset reveals distinct cross-linguistic patterns, with Romanian reviews averaging 2.4× longer than Italian reviews, underscoring the importance of accounting for linguistic variation in multilingual NLP systems. Future work should explore incorporating the unlabeled data through semi-supervised learning and extend the framework to additional low-resource languages.

Ethics Statement

The RoIt-XMASA dataset was compiled from publicly available e-commerce reviews, in accordance with the terms of service of the source platforms. To protect user privacy, all personally identifiable information and author metadata were removed, leaving only review text, titles, and ratings. The dataset is released for academic research purposes only. Although rigorous cleaning was applied, the reviews were not filtered for toxic language to preserve natural linguistic distributions; consequently, the corpus may contain offensive content, typical of unmoderated user-generated text.

Acknowledgements

This work was supported by the National University of Science and Technology POLITEHNICA Bucharest through the PubArt program.

References

Araujo, M.; Reis, J.; Pereira, A.; and Benevenuto, F. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st annual ACM symposium on applied computing*, 1140–1145.

Augustyniak, L.; Woźniak, S.; Gruza, M.; Gramacki, P.; Rajda, K.; Morzy, M.; and Kajdanowicz, T. 2023. Massively multilingual corpus of sentiment datasets and multi-faceted

sentiment classification benchmark. *Advances in Neural Information Processing Systems*, 36: 38586–38610.

Avram, A.-M.; Bănescu, E.-I.; Robea, A.-T.; Cercel, D.-C.; and Cercel, M.-C. 2025a. MoRoVoc: A Large Dataset for Geographical Variation Identification of the Spoken Romanian Language. *arXiv preprint arXiv:2509.16781*.

Avram, A.-M.; Iuga, A.; Manolache, G.-V.; Matei, V.-C.; Micliuş, R.-G.; Muntean, V.-A.; Sorlescu, M.-P.; Şerban, D.-A.; Urse, A.-D.; Păiş, V.; et al. 2024. Histnero: Historical named entity recognition for the romanian language. In *International Conference on Document Analysis and Recognition*, 126–144. Springer.

Avram, A.-M.; Timpuriu, M.; Iuga, A.; Matei, V.-C.; Taiatu, I.-M.; Găină, T.; Cercel, D.-C.; Cercel, M.-C.; and Pop, F. 2025b. RoLargeSum: A Large Dialect-Aware Romanian News Dataset for Summary, Headline, and Keyword Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2049–2066.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 440–447.

Chen, X.; Sun, Y.; Athiwaratkun, B.; Cardie, C.; and Weinberger, K. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6: 557–570.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Prettenhofer, P.; and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 1118–1127.

Qi, Y.; and Shabrina, Z. 2023. Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. *Social network analysis and mining*, 13(1): 31.

Rajda, K.; Augustyniak, L.; Gramacki, P.; Gruza, M.; Woźniak, S.; and Kajdanowicz, T. 2022. Assessment of Massively Multilingual Sentiment Classifiers. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 125–140.

Saad-Falcon, J.; Khattab, O.; Santhanam, K.; Florian, R.; Franz, M.; Roukos, S.; Sil, A.; Sultan, M.; and Potts, C. 2023. UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 11265–11279.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vettoruzzo, A.; Bouguelia, M.-R.; Vanschoren, J.; Rögnvaldsson, T.; and Santosh, K. 2024. Advances and challenges in meta-learning: A technical review. *IEEE transactions on pattern analysis and machine intelligence*, 46(7): 4763–4779.

Wan, X. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 235–243.

Winata, G. I.; Aji, A. F.; Cahyawijaya, S.; Mahendra, R.; Koto, F.; Romadhony, A.; Kurniawan, K.; Moeljadi, D.; Prasojo, R. E.; Fung, P.; et al. 2023. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834.

Xing, B.; Ying, X.; Wang, R.; and Guo, R. 2025. Multimodal Prompt Alignment with Fine-grained LLM Knowledge for Unsupervised Domain Adaptation. *International Journal of Computer Vision*, 1–22.

Xu, K.; and Wan, X. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 511–520.

Ye, H.; Tan, Q.; He, R.; Li, J.; Ng, H. T.; and Bing, L. 2020. Feature Adaptation of Pre-Trained Language Models across Languages and Domains with Robust Self-Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7386–7399.

Zhang, B.; Tian, Y.; Wang, S.; Tu, Z.; Chu, D.; and Shen, Z. 2024. GongBu: Easily Fine-tuning LLMs for Domain-specific Adaptation. In *Proceedings of the 33rd ACM Inter-*

national Conference on Information and Knowledge Management, 5309–5313.

Zhao, C.; Wu, M.; Yang, X.; Zhang, W.; Zhang, S.; Wang, S.; and Li, D. 2024. A systematic review of cross-lingual sentiment analysis: Tasks, strategies, and prospects. *ACM Computing Surveys*, 56(7): 1–37.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this work introduces a multilingual dataset to improve sentiment analysis for underrepresented languages such as Romanian, following established ethical scraping and cleaning protocols.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the contributions regarding the RoIt-XMASA dataset and the multi-target adversarial training framework are outlined in the abstract and introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we describe the multi-adversarial objective and meta-learning strategy used to balance sentiment discrimination with domain/language invariance.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss the rating distribution and textual characteristics, noting peaks at the extreme ratings and differences in review styles between Romanian and Italian.
   (e) Did you describe the limitations of your work? Yes, limitations related to resource scarcity for Romanian and the challenges of the double shift in the domain and language are discussed in the Introduction.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, the paper mentions considerations regarding cultural nuances and potential model failures in capturing these subtleties.
   (g) Did you discuss any potential misuse of your work? No, because the primary application is sentiment analysis for e-commerce reviews, which has low potential for harmful misuse beyond standard concerns for all NLP models.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we detail rigorous cleaning, language verification, manual verification for quality, and provide a HuggingFace link for responsible data release.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, the data collection followed native linguistic patterns and language-specific acquisition strategies to ensure authentic and ethical data usage.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the dataset is released on HuggingFace (see Footnote 1).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we specify splits in Table 1 and provide full hyperparameter details in Appendices.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because single-run results were reported for benchmarking; however, a fixed random seed (42) was used for reproducibility.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because the focus was on F1-score performance and methodological novelty, though mixed-precision training was noted to reduce memory.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we evaluate across three domains and two languages using both encoder-based and LLM baselines to validate our adversarial framework.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, as the task is four-class sentiment analysis, where errors represent deviations in user perception rather than critical system failure.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we cite the original XMASA dataset creators and the developers of models like XLM-R and Llama-3.1.
   (b) Did you mention the license of the assets? No, because the models used (Llama, XLM-R) are under well-known open-source licenses cited in the references.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the RoIt-XMASA dataset is available via the provided HuggingFace link.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, data was collected from public e-commerce platforms using language-specific acquisition strategies.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, manual verification by native speakers assessed text quality and alignment, ensuring samples represent unique review instances.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? No, though the dataset is publicly hosted on HuggingFace to ensure accessibility.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? No, but detailed statistical characteristics are provided in Section 3 and Appendices.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? N/A
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A
   (d) Did you discuss how data is stored, shared, and deidentified? Yes, we applied cleaning and normalization to ensure data quality and integrity.

Cross-Language Analysis

The RoIt-XMASA dataset reveals differences in the reviewing patterns between Italian and Romanian users. Romanian reviews are, on average, 2.4 times longer than Italian reviews (89.02 vs. 37.34 tokens), suggesting more elaborate review styles in Romanian online communities. This pattern is clearly visible in Figure 3, where the Romanian distribution shows a longer tail with reviews that frequently exceed 500 tokens, while Italian reviews concentrate below 200 tokens.

Title usage patterns show an inverse relationship, as illustrated in Figure 4. Italian reviews show a strong preference for including titles, whereas Romanian reviews rarely include titles. The Italian distribution shows clear peaks at 2-5 tokens, suggesting standardized title formats, while the Romanian distribution is heavily concentrated at zero, indicating different platform conventions or user behaviors.

These language-specific patterns in RoIt-XMASA have important implications for model architecture choices, particularly regarding maximum sequence length and attention mechanisms. The substantial difference in review lengths suggests that models may benefit from language-specific pre-processing or architectural adaptations.

Cross-Domain Analysis

Domain-specific review patterns emerge clearly from the RoIt-XMASA token distributions shown in Figure 5. Book reviews have the highest mean length (76.27 tokens), with a broad distribution that extends beyond 1,000 tokens, reflecting detailed literary analysis and plot discussion. Music reviews follow closely (71.53 tokens) with similar variance, while movie reviews are notably more concise (41.73 tokens) with a distribution concentrated below 500 tokens.

The usage of titles varies by domain, as shown in Figure 6. Music reviews exhibit the highest variance in title length (std: 11.12), with a bimodal distribution showing peaks at both 0 and 50+ tokens, reflecting users who either omit titles entirely or include full album/track listings. Book and movie reviews show more consistent usage patterns of titles with distributions concentrated below 20 tokens.

Interesting interaction effects emerge when examining language-domain combinations in RoIt-XMASA. The Italian-Romanian length ratio ranges from 1.6:1 in movie reviews to 3.7:1 in music reviews, indicating that cultural factors influence discussions of different media types in each language. These substantial cross-domain differences validate the continued relevance of the domain adaptation challenges identified by Blitzer, Dredze, and Pereira (2007), now extended to multilingual settings through the RoIt-XMASA dataset.

Text Normalization Pipeline

The following normalization rules were applied to all reviews in the RoIt-XMASA dataset to ensure consistent text representation for transformer-based models (Vaswani et al. 2017) while preserving essential, semantic, and syntactic information for sentiment analysis:

1. **Punctuation normalization**: Multiple consecutive punctuation marks were reduced to a maximum of three instances (e.g., "!!!!!" → "!!!"), preserving emphasis while preventing excessive repetition. This includes ellipses ("....." → "...") and mixed punctuation patterns.
2. **Whitespace standardization**: All sequences of multiple whitespace characters (spaces, tabs, and non-breaking spaces) were replaced with single spaces, while leading and trailing whitespaces were removed from each review.
3. **URL and email handling**: URLs and email addresses were replaced with special tokens `[URL]` and

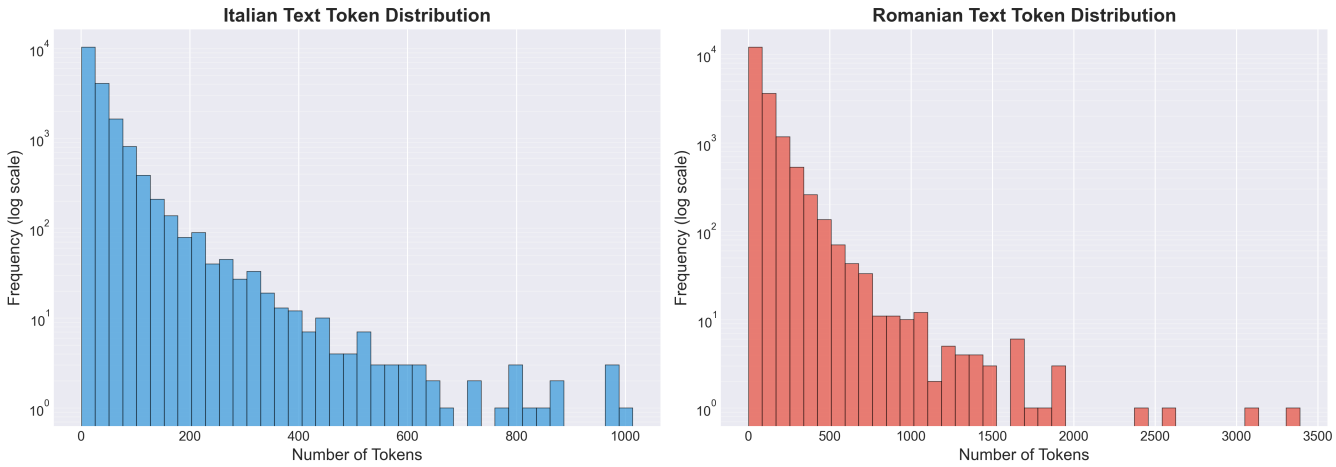

Figure 3: Token distribution of the text in RoIt-XMASA, grouped by language.

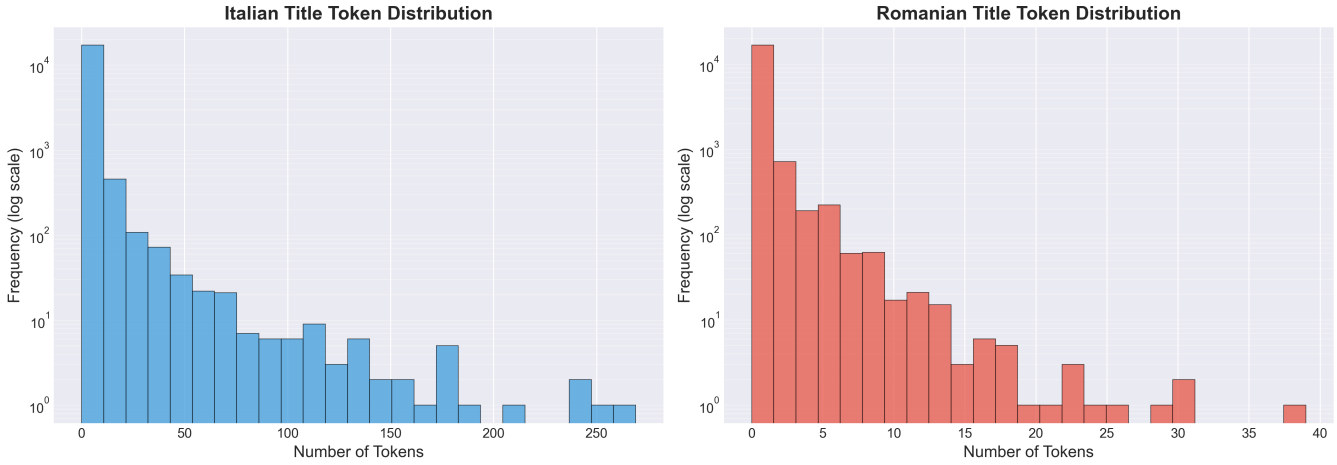

Figure 4: Token distribution of the title in RoIt-XMASA, grouped by language.

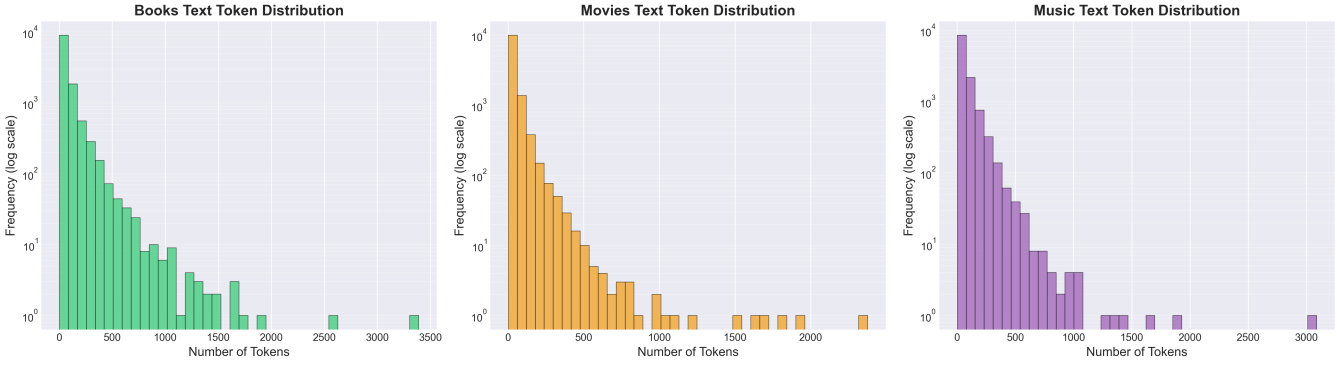

Figure 5: Token distribution of the text in RoIt-XMASA, grouped by domain.

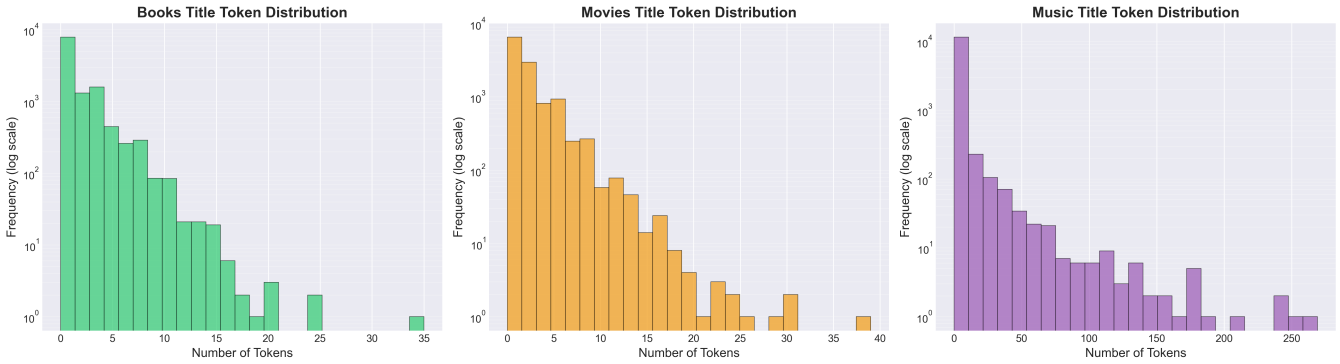

Figure 6: Token distribution of the title in RoIt-XMASA, grouped by domain.

`[EMAIL]`, respectively, preserving the information that external references existed, while removing potentially noisy string patterns.

These normalization steps were designed to balance data cleaning with information preservation, ensuring optimal performance for transformer-based models while maintaining the linguistic characteristics essential for sentiment analysis in Italian and Romanian.

Hyperparameters for Multilingual Encoder Models

Table 5 outlines the hyperparameters employed when fine-tuning M-BERT, XLM, and XLM-R for sentiment classification tasks. We employed early stopping with a patience of 3 epochs based on the validation set F1-score to prevent overfitting. The meta-learning rate η for the adversarial coefficient updates was set to 0.01, with the coefficients initialized to $\lambda_1 = \lambda_2 = 0.5$ and reduced in the range $[0, 2]$ to ensure training stability. The adversarial coefficients were updated every 100 training steps using the meta-learning procedure described in Section 4.1.

All experiments used mixed-precision training (FP16) to reduce memory consumption and accelerate training. We set the random seed to 42 for reproducibility across all runs. The classification head consisted of a single linear layer with dropout applied before the final projection. For the adversarial discriminators (i.e., domain and language), we used identical single-layer architectures with the same dropout rate.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	32
Training Epochs	3–5
Max Sequence Length	128
Dropout Probability	0.1
Weight Decay	0.01
Warmup Ratio	0.1
Gradient Clipping	1.0
Meta Learning Rate (η)	0.01
Early Stopping Patience	3
Random Seed	42

Table 5: Hyperparameters used for fine-tuning multilingual encoder models.

Hyperparameters for LLMs

This section details the configuration for the LLM baselines. During the evaluation phase, we strictly employed greedy decoding (setting the temperature to 0.0) to ensure deterministic and reproducible sentiment rating predictions. The inference process used a maximum limit of 5 for new tokens, which was sufficient to generate a single-number rating.

The unsupervised domain adaptation phase focused on the 202,141 unlabeled reviews provided in the RoIt-XMASA dataset. This stage used the AdamW optimizer (Loshchilov and Hutter 2019) with a cosine learning rate scheduler. The value-based hyperparameters for the LoRA adaptation are provided in Table 6.

Hyperparameter	Value
Epochs	1
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Learning Rate	2e5
Warmup Ratio	0.05
Weight Decay	0.01
Global Batch Size	16

Table 6: Hyperparameter settings for the LLM unsupervised adaptation and inference phases.

LLM Evaluation Prompts

In this section, we present the templates used for the zero-shot and few-shot evaluations of the LLMs. The templates provided below are the translated versions used for querying the models (i.e., from Romanian and Italian); for the few-shot configurations, the placeholders for `Title`, `Review`, and `Rating` were populated with the respective language-specific samples from the RoIt-XMASA dataset.

Zero-Shot

You are a review rating predictor. Given a review text, predict its rating on a scale of 1 to 5 (except 3).

1 = Very negative
2 = Negative
4 = Positive
5 = Very positive

Only respond with a single number (1, 2, 4, or 5).
Title: {}
Review: {}
Rating:

Multi-Shot

You are a review rating predictor. Given a review text, predict its rating on a scale of 1 to 5 (except 3).

1 = Very negative
2 = Negative
4 = Positive
5 = Very positive

Only respond with a single number (1, 2, 4, or 5).

Here are some examples:
Title1: {}
Review1: {}
Rating1: {}
...

Now predict the rating for this review:
Title: {}
Review: {}
Rating: