MALTO at FadeIT: A BERT-Based System for Multi-Label Fallacy Detection in Italian Social Media

Matin Salami[1], Luca M. Rodia[1], Vladimir Schiau[1], Evren A. Munis[1], Claudio Savelli[1] and Flavio Giobergia[1]

[1] *Politecnico di Torino, Turin, Italy*

Abstract

Fallacies in social media posts are common and can make texts less trustworthy, and more likely to foster biased views. In this work, we study multi-label fallacy classification in Italian social media posts by detecting 20 different fallacy types that can appear in the same post. We address this task using encoder-based language models. In particular, we build a fallacy classifier based on AlBERTo, a BERT-style encoder for Italian, and use a global multi-label threshold to turn the model's outputs into the final set of fallacy labels. Our results indicate that paraphrase-based augmentation can degrade performance in fallacy detection, likely by introducing label noise, while the choice of loss function substantially affects the trade-off between micro- and macro-averaged performance.

Keywords

fallacy detection, multi label classification, encoder based models, deep learning, natural language processing

1. Introduction

Fallacious reasoning is pervasive in online public discourse, where emotionally charged and fast-paced interactions often foster misleading argumentative strategies [1, 2]. Detecting such fallacies is crucial for supporting critical thinking, promoting healthier public debate, and mitigating the spread of manipulative or harmful content [3]. However, this task is particularly challenging due to the subjective nature of the interpretation of the fallacy and the resulting disagreement among human annotators.

The FadeIT [4] (Fallacy Detection in Italian Social Media Texts) shared task, proposed within the EVALITA 2026 [5] campaign, addresses this problem for the Italian language by focusing on fallacy detection in social media posts over a four-year time span [4]. This paper focuses on Subtask A, which targets coarse-grained, post-level fallacy detection. The task is formulated as a multi-label classification problem, where each post may contain zero, one, or multiple fallacies, as annotated by two annotators. This setting reflects real-world conditions, where argumentative interpretation is often ambiguous rather than governed by a single ground truth.

In this work[1], we propose a classifier based on encoder-only transformer-based models. In the submitted results, we used the AlBERTo model [6], which encodes each social media post into token-level representations. These embeddings are aggregated via mean pooling to obtain a fixed-size representation of the post, which is then fed to a lightweight classification head for multi-label prediction.

Our best approach, described above, achieves a micro-averaged F1 score of 0.54 on the official test set for Subtask A. Using different configurations with augmented data and weighted loss resulted in a score of 0.374, while an additional approach incorporating description awareness reached 0.466 on the test set.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

s306400@studenti.polito.it (M. Salami); s338250@studenti.polito.it (L. M. Rodia); s353706@studenti.polito.it (V. Schiau); evrenayberk.munis@studenti.polito.it (E. A. Munis); claudio.savelli@polito.it (C. Savelli); flavio.giobergia@polito.it (F. Giobergia)

0000-0002-0877-7063 (C. Savelli); 0000-0001-8806-7979 (F. Giobergia)

[1]The code to replicate the experiments can be found at https://github.com/MAL-TO/FADEIT-2026

2. Related Works

This challenge addresses two main obstacles: handling an unbalanced dataset and extracting meaningful information from unstructured text. The following subsections review the main approaches found in the literature to solve these problems

Multilabel Classification: A method proposed by Lanchantin et al. [7] adapts transformer architectures for multilabel classification. In this framework, the image features are processed alongside label embeddings, which are randomly masked during training to predict the full label set. This strategy effectively captures label correlations while maintaining robustness at inference time; by training on instances where all labels are masked, the model also learns to perform accurate classification solely from visual data when no partial label information is available.

A sequence-to-sequence approach is proposed in Yang et al. [8], where the authors apply an attention mechanism to adaptively select the most informative words for each predicted label. This model treats multi-label classification as a generation problem, capturing high-order correlations by conditioning the prediction of the next label on the previously generated ones. Furthermore, they introduce a Global Embedding mechanism: this combines the embedding of the top-predicted label with a probability-weighted average of the entire label distribution from the previous time step.

Finally, Da San Martino et al. [9] examines propaganda techniques and logical fallacies in news articles, and proposes a Multi-Granularity Network. In this architecture, a sentence-level classification drives a 'Gate' mechanism: if a sentence is predicted as non-propagandistic, the gate suppresses further fragment-level predictions, thereby significantly reducing false positives.

Multilabel Classification under Class Imbalance: A major challenge in fallacy classification is the highly skewed distribution of labels. For a better description go to data distribution. Existing approaches mainly address this problem through loss re-weighting strategies and data-level balancing techniques. Loss-based methods adapt the standard cross-entropy to weight samples differently based on their ground-truth labels, aiming to reduce the influence of frequent classes. Focal Loss [10] achieves this by down-weighting easy examples and emphasizing harder instances, effectively focusing on the model's certainty. Similarly, Class-Balanced Loss [11] introduces the notion of effective information, observing that the marginal information provided by each additional example decreases as class frequency increases. The loss is reduced proportionally to the class frequency, allowing the model to focus on underrepresented samples.

Data Augmentation via LLM An alternative strategy for addressing long-tailed label distributions is data augmentation via oversampling, which artificially rebalances the dataset by increasing the number of instances for minority classes [12]. A representative method in this line of research is that of He et al. [13], in which the authors generate synthetic samples in proportion to a balancing coefficient $\beta \in [0, 1]$. In this framework, $\beta = 1$ corresponds to a fully balanced dataset, while $\beta = 0$ reduces to the non-augmented data distribution. When the augmented data are used to train a decision tree classifier, the authors report consistent improvements over both the non-augmented setting and the SMOTE approach [14], achieving approximate recall gains of +8% and +32% respectively. More recent approaches to data augmentation for logical fallacy detection rely on large language models (LLMs) to generate synthetic argumentative content. This direction is explored by Poliakov and Shvai [15], who employ a Retrieval-Augmented Generation (RAG) framework to condition an LLM on scientific literature, enabling the generation of contextually grounded, plausible, and fallacious arguments. A related approach is presented in Mouchel et al. [16], where GPT-3.5-Turbo is used to generate multiple logical arguments from the EXPLAGRAPHS dataset [17]. Furthermore, Glazkova and Zakharova [18] demonstrate that prompt-based LLM augmentation strategies can yield greater performance compared to models trained only on original data. Following this line of work, we tested ChatGPT to generate paraphrases of selected training samples, which are then added to the training data before model optimization. The paraphrases are generated exclusively from the training set, and the original labels

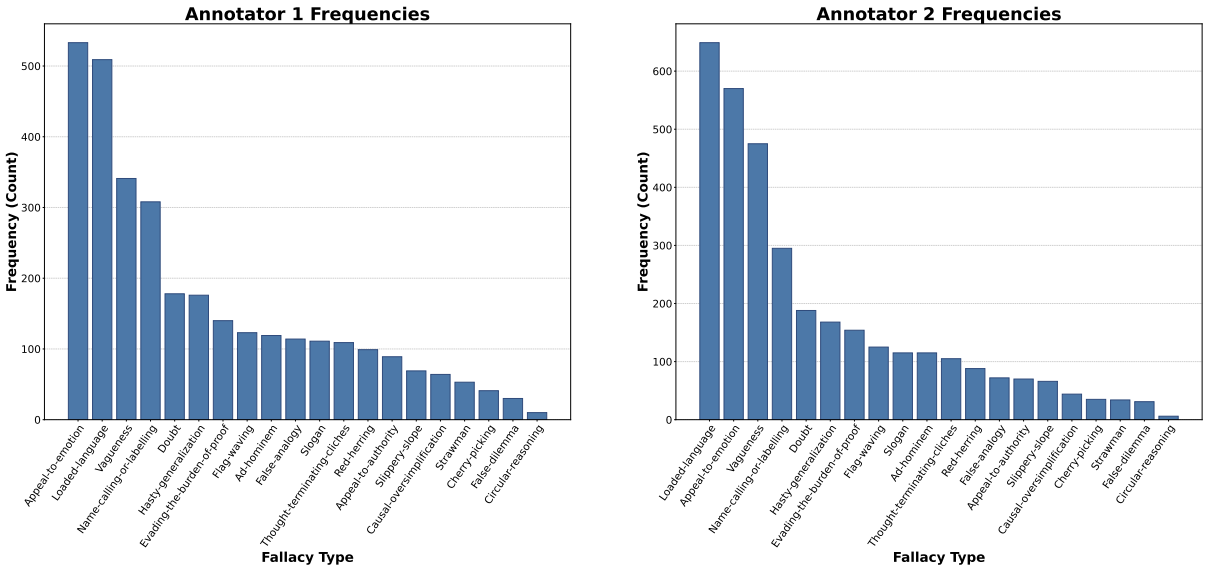

Figure 1: Label counts per annotator across the complete dataset.

are retained for all augmented instances. After augmentation, the dataset remains heavily imbalanced, so we also address this by using different weighted loss functions. The exact prompt used for augmentation is reported in appendix B for reproducibility.

Fallacy detection: A central challenge in text classification is extracting meaningful semantic representations that capture the relevant concepts expressed in the text. This is further amplified in tasks such as logical fallacy detection, where labels often correspond to abstract rhetorical patterns rather than surface-level lexical clues. An effective approach is proposed in Wang et al. [19], where both the input text and the label descriptions are encoded into dense vector representations. Each text and label is mapped to a single embedding, and the resulting representations are subsequently aggregated and fed into a neural classifier. In contrast, Çöltekin et al. [20] explores a more traditional machine learning approach based on Support Vector Machines (SVM) trained on word n-gram features. Despite its relative simplicity, this method remains competitive due to its computational efficiency and interpretability. Moreover, the reduced training cost enables extensive hyperparameter tuning, resulting in reasonable performance even without deep contextual representations.

3. Challenge Description

Task: The objective of subtask A is to perform multi-label classification at the post level. This means that, given a social media post, the model must predict, for both annotators, the set of fallacy types present in the text. The model performance is evaluated using the micro-averaged F1 score, which is the metric used for the main evaluation and submission ranking.
Data: The FadeIT shared task is based on FAINA [4], a dataset for fallacy detection in Italian social media posts. The data, spanning a four-year time period, cover public discourse on three socially relevant topics: migration, climate change, and public health. Posts are anonymized by replacing user mentions and sensitive information with placeholders such as [USER], [URL], [EMAIL], and [PHONE]. The dataset includes 20 fallacy types and provides annotations from two independent annotators, allowing multiple fallacy labels per post and capturing genuine disagreement in fallacy interpretation. The data are split into a train set and a test set. The train set consists of 1152 posts, while the test set contains 288 posts.
Data distribution: Several fallacies represent only a small fraction of the dataset, such as circular reasoning and false dilemma, which combined account for less than 1resulting in a long-tailed label

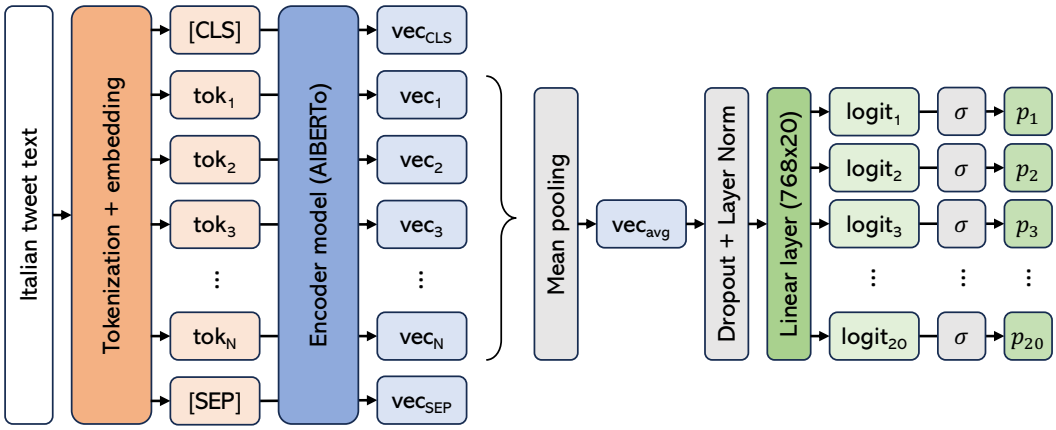

Figure 2: Architecture of the proposed multi-label text classification model. An Italian tweet is tokenized and encoded using AlBERTo. Contextualized token representations are aggregated via attention-mask-aware mean pooling, followed by dropout, layer normalization, and a fully connected classification layer that produces one logit per label. Sigmoid activations are applied during inference to obtain label probabilities.

distribution. This issue is exacerbated in a multilabel setting, where minority labels often co-occur with more frequent ones.

4. Methodology

Our approach is inspired by previous methods for multi-label text classification introduced by Fallah et al. [21]. It builds upon: 1) a BERT-adapted model and 2) a thresholding technique. The two main parts of it are explained in detail in the next two sections.

4.1. BERT Adaptation

Transformer-based pretrained language models such as BERT and its variants have become a dominant approach and have pushed the state of the art across many NLP tasks [22, 23]. Since our dataset consists of Italian tweets from social media, we employ AlBERTo, a BERT-based language model pretrained on approximately 200 million Italian tweets [6]. We further include UmBERTo, another monolingual Italian RoBERTa-based model [24], and XLM-RoBERTa-base, a multilingual transformer [25], as evaluated architectures for the proposed pipeline.

Each text instance in the dataset is annotated according to two distinct professional annotators [26]. Rather than learning separate predictors for the two schemes, we merge the two label sets into a unified label space and train a single multi-label classifier that predicts one shared set of labels.

In our approach, instead of relying on the special `[CLS]` token, the encoder produces a sequence of contextualized token representations $H = \{h_1, h_2, \ldots, h_T\}$ for an input sequence of length T, where each $h_t \in \mathbb{R}^d$ and d denotes the hidden size of the encoder. To obtain a fixed-size representation of the entire input text, we apply mean pooling over the token dimension:

$$h = \frac{1}{T} \sum_{t=1}^{T} h_t$$

Mean pooling is used to obtain a document-level representation, as global pooling methods (e.g., average pooling) have been found to be particularly effective for classification tasks [27].

The resulting document representation is passed through a dropout layer for regularization and a layer normalization step [28] to improve training stability. Finally, a fully connected linear layer maps the normalized representation to a vector of K output logits, one for each label. The model architecture is depicted in Figure 2.

4.2. Global Thresholding Strategy

Since the task is formulated as multi-label classification, the model outputs are converted into binary predictions using a decision threshold. Instead of fixing this threshold a priori, we adopt a global threshold selection strategy. Specifically, a global classification threshold $\tau \in [0, 1]$ is selected by maximizing the micro-averaged F1 score on the validation set. The threshold is varied from 0 to 1 with a step size of 10^{-1}, and the value yielding the best validation performance is selected. The optimal threshold is then applied to the test set for inference which in our case is 0.5.

4.3. Label Description–Aware Architecture

In addition to the standard setup, we explored a label-aware variant of the BERT-based classifier aimed at explicitly incorporating semantic information from the textual descriptions of fallacy labels. The main motivation was to move beyond a purely document-level classification paradigm and instead model fallacy detection as a semantic compatibility problem between the input text and the conceptual description of each label. In this architecture, each post is encoded using a pretrained Italian BERT model, and the [CLS] token representation is used as the document embedding. Each fallacy label is represented by a learnable embedding initialized from its textual description. For each (text, label) pair, the projected document embedding is concatenated with the corresponding label embedding and fed into a shared feed-forward network that outputs a single binary logit. This process is repeated independently for all labels, producing multi-label predictions for each post. The model is trained using Focal Loss to address class imbalance, and class-specific decision thresholds are selected on the validation set to optimize F1-score. This formulation allows the model to explicitly encode relationships between textual content and fallacy definitions, encouraging semantic alignment.

5. Experimental Setup

This section describes the experimental protocol adopted to evaluate the proposed model and the baseline methods. For all baseline techniques, we maintain the original model implementations to ensure a fair comparison.

5.1. Training Procedure

All models are optimized using AdamW [29] with a learning rate of 2×10^{-5} and trained for 10 epochs. We employ a cosine learning rate schedule with linear warm-up, where the warm-up phase corresponds to the first 10% of the total training steps. Performance is evaluated using micro-averaged precision, recall, and F1-score. All experiments are conducted on a machine equipped with an AMD Ryzen 9 7940HX CPU and an NVIDIA GeForce RTX 4070 GPU. The full configuration table is in appendix A.

To obtain training, validation, and test splits that preserve label imbalance and label co-occurrence patterns, we adopt an iterative stratification strategy. This procedure mitigates the risk of rare labels being absent from the validation or test sets due to random splitting. We partition the data into three subsets: a training set containing 70% of the instances, and validation and test sets each containing 15% of the data.

5.2. Loss Function Configurations

To address class imbalance, we experimented with three loss functions: (i) Binary Cross-Entropy, (ii) Binary Cross-Entropy with positive class weighting, and (iii) Focal Loss.

Binary Cross-Entropy. BCE is a standard choice for multi-label classification, treating each label as an independent binary prediction. This formulation assumes a balanced label distribution and assigns equal importance to all classes, which can be suboptimal in highly imbalanced settings.

Binary Cross-Entropy with Positive Class Weights. To mitigate label imbalance, we employed a weighted variant of BCE that assigns higher importance to positive instances of rare labels. For

Table 1
Performance comparison across experimental settings. **Test set** indicates the evaluation split: *Ours* refers to the internally constructed test set, while *Official* refers to the competition's official test set.

Method	Test set	Micro-average			Macro-average		
		Precision	Recall	F1	Precision	Recall	F1
Baseline	Official	0.385	0.142	0.208	0.151	0.034	0.051
AlBERTo (BCE with O-data)	Official	**0.557**	**0.536**	**0.546**	**0.419**	**0.309**	**0.326**
AlBERTo (BCE(weights) with Aug-data)	Official	0.374	0.45	0.408	0.235	0.30	0.256
AlBERTo (description-aware with Aug-data)	Official	0.466	0.437	0.450	0.281	0.233	0.232
AlBERTo (BCE with O-data)	Ours	0.574	0.576	**0.575**	**0.36**	0.315	0.312
AlBERTo (BCE(weights) with O-data)	Ours	0.436	**0.622**	0.513	0.318	**0.4652**	**0.37**
AlBERTo (Focal with O-data)	Ours	**0.621**	0.463	0.53	0.349	0.247	0.276
AlBERTo (BCE with Aug-data)	Ours	0.45	0.495	0.471	0.25	0.244	0.2316
AlBERTo (BCE(weights) with Aug-data)	Ours	0.38	0.40	0.43	0.245	0.323	0.27
AlBERTo (Focal loss with Aug-data)	Ours	0.38	0.493	0.43	0.245	0.323	0.27
UmBERTo (BCE with O-data)	Ours	0.5964	0.5539	0.574	0.298	0.273	0.269
XLM-Roberta-Base (BCE with O-data)	Ours	0.613	0.528	0.567	0.288	0.263	0.263

each label i, a positive weight w_i is computed as the ratio between negative and positive samples, $w_i = N_i^- / (N_i^+ + \epsilon)$, where N_i^+ and N_i^- denote the number of positive and negative occurrences of label i, respectively. This strategy increases the contribution of rare labels during training and has been shown to improve robustness in imbalanced multi-label classification settings [30]. In our experiments, the weights were computed directly from the training data.

Focal Loss. The metric extends BCE by down-weighting well-classified examples and focusing training on harder instances. Starting from the per-label BCE loss ℓ_i^{BCE}, we define the probability assigned to the ground-truth class as $p_{t,i} = \exp(-\ell_i^{\text{BCE}})$. The focal loss then applies a modulating factor $(1 - p_{t,i})^\gamma$ and an α-balancing term to reweight positive and negative targets, yielding:

$$\mathcal{L}_{\text{FL}} = \sum_{i=1}^{C} \alpha_{t,i} (1 - p_{t,i})^\gamma \ell_i^{\text{BCE}} \tag{1}$$

where $\gamma \geq 0$ controls the focus on hard examples. This objective was originally introduced for dense object detection [10] and has since been adopted in imbalanced multi-label learning.

5.3. Data Augmentation with ChatGPT

Following prior work on LLM-based data augmentation, we leveraged ChatGPT to generate paraphrases of selected training instances, which were subsequently added to the training set prior to model optimization. Importantly, paraphrases were generated exclusively from the training data to prevent data leakage, and the original labels were preserved for all augmented samples. To specifically target class imbalance, we generated two paraphrases for each instance belonging to fallacy labels with fewer than 100 training examples. This process resulted in 2,445 additional training instances. Despite this augmentation strategy, the resulting dataset remained heavily imbalanced. To further mitigate this issue, we combined data augmentation with weighted loss functions, as described in the previous subsection. For reproducibility, the exact prompt used for paraphrase generation is provided in Appendix B.

6. Results

We conducted eight different experiments using two data configurations (original and augmented) and three loss functions. Due to the competition constraints, only two model configurations were

Table 2
Label distribution across annotation sets (A1, A2) and per-label F1 scores. Results are obtained with AlBERTo model with normal BCE loss and without data augmentation. Cell colors introduced as a visual aid to highlight higher (green) and lower values (red), for each column.

Code	Label	freq in A1	freq in A2	F1 (%)
AH	Ad hominem	119	115	50.86
AA	Appeal to authority	89	70	11.49
AE	Appeal to emotion	533	570	71.06
CO	Causal oversimplification	64	44	6.25
CP	Cherry picking	41	35	12.50
CR	Circular reasoning	10	6	0.00
DO	Doubt	178	188	70.32
EP	Evading the burden of proof	140	154	34.45
FA	False analogy	114	72	15.84
FD	False dilemma	30	31	0.00
FW	Flag waving	123	125	42.45
HG	Hasty generalization	176	168	33.51
LL	Loaded language	509	649	68.03
NC	Name calling or labelling	308	295	59.62
RH	Red herring	99	88	16.11
SS	Slippery slope	69	66	20.00
SL	Slogan	111	115	55.32
ST	Strawman	53	34	0.00
TC	Thought-terminating cliché	109	105	26.58
VA	Vagueness	341	475	58.23

evaluated on the official test set provided by the organizers. Overall, our experiments demonstrate the effectiveness of BERT-based models for multi-label classification in fallacy detection tasks.

Table 1 presents the results obtained under the different experimental setups. In both submissions evaluated on the official test set, our models outperform the competition's baseline. The most notable result is achieved with the BCE loss without class weighting or data augmentation, yielding the best overall performance on the official test set: a micro-F1 score of 0.546 and a macro-F1 score of 0.326.

The second submitted model, trained using BCE loss with POS-weighting on augmented data, also outperforms the baseline across all evaluation metrics. However, its performance remains below that of the best performing configuration. These results indicate that, while class reweighting and data augmentation can provide benefits, they do not necessarily guarantee superior performance compared to training on the original data alone.

The description-aware model, trained using Focal loss on augmented data, achieved a micro-F1 score of 0.4505 on the official test set. Despite its more structured design, the results suggest that explicitly modeling label semantics does not necessarily achieve better overall performance.

Despite the clear improvements over the baseline, all submitted models remain weak in identifying fallacy labels with very low frequency, as well as fallacies whose detection depends primarily on argumentative structure rather than lexical or semantic content, such as circular reasoning and Causal oversimplification, as shown in Table 2. This indicates that rare labels and structure-driven fallacies continue to pose a significant challenge for models that rely mainly on contextual text representations.

6.1. Evaluation of different loss functions

Due to the submission constraints of the shared task, it was not possible to evaluate all loss function configurations on the official test set. Therefore, in this section, we compare different loss functions using our internally constructed test set and analyze their impact on model performance with respect to the task evaluation metrics.

We experiment with three loss formulations: BCE, weighted BCE, and Focal Loss. As reported in Table 1, all configurations substantially outperform the baseline system, regardless of the loss function

employed. Overall, the best micro-averaged F1 scores are achieved by models trained with standard BCE loss across all architectures, namely AlBERTo, UmBERTo, and XLM-RoBERTa-base, which obtain micro-F1 scores of 0.575, 0.574, and 0.567, respectively.

This behavior can be explained by the properties of the BCE loss, which assigns equal importance to all labels and therefore encourages the model to maximize the total number of correctly predicted labels, irrespective of their frequency. Given that the task's primary evaluation metric, micro-F1, is dominated by frequent labels, this loss formulation aligns well with the evaluation objective.

In contrast, BCE with class weights and Focal Loss explicitly emphasize rare labels by increasing their contribution to the training objective. As a consequence, these loss functions generally yield higher macro-averaged F1 scores, reflecting improved performance on minority classes, but at the cost of lower micro-F1 scores. This highlights a clear trade-off between optimizing overall label-wise accuracy and improving sensitivity to infrequent fallacy types.

Given the task evaluation protocol, we therefore adopt BCE as the primary loss function in our final models.

6.2. Use of augmented data and limitations

The results reported in Table 1 indicate that augmenting the training data with synthetic paraphrases does not lead to an improvement in overall model performance. We identify two main factors that likely contribute to this outcome. First, although LLM-based paraphrasing increases the number of training instances, it can attenuate fallacy-specific rhetorical cues that are characteristic of social media discourse, such as framing strategies, slang, irony, and informal stylistic markers. As a consequence, inheriting the original labels for paraphrased instances may introduce both distribution shift and label noise. Second, fallacy detection is inherently sensitive to pragmatic and stylistic features that go beyond literal semantic equivalence. Even when a paraphrase preserves the core meaning of the original text, subtle changes in tone, emphasis, or rhetorical structure may alter the underlying fallacy signal. Together, these factors provide a plausible explanation for the lack of performance gains observed with paraphrase-based data augmentation in our experiments.

Despite these limitations, the results obtained by our best-performing model are promising, as they consistently outperform the baseline systems and other methods reported by Ramponi and Tonelli [26]. Table 2 reports the average F1 score for each individual fallacy label. As shown, the model achieves strong performance on frequent labels, such as *Appeal to emotion* (average count = 551, F1 = 71.06) and *Doubt* (average count = 183, F1 = 70.32). In contrast, performance drops substantially for less frequent fallacies, including *Red herring* (average count = 93, F1 = 16.11) and *False analogy* (average count = 93, F1 = 15.84). Moreover, the model fails to correctly identify three fallacy types, namely *Circular reasoning*, *False dilemma*, and *Strawman*.

These results highlight the intrinsic difficulty of fallacy detection, which requires models to reason beyond surface-level lexical cues. While BERT-based architectures have demonstrated strong performance across a wide range of text-classification tasks, fallacy detection often depends on higher-level discourse structure and argumentative patterns. For example, *Circular reasoning* fallacy requires understanding how claims recursively support themselves within the same text, rather than relying on isolated word-level features. As discussed by Ramponi et al. [4], successfully detecting such fallacies requires modeling global textual structure and reasoning over the full argument.

7. Conclusion

This work presents a system developed for the FadeIT task at EVALITA, which addresses multi-label fallacy classification in Italian social media posts. Our approach fine-tunes a BERT-based language model using a global multi-label classification threshold and evaluates the impact of alternative loss functions and data augmentation strategies. Experimental results indicate that models trained with paraphrase-based augmentation tend to underperform, suggesting that automatically generated paraphrases may introduce noise and distort fallacy cues present in the original data. Among the loss formulations,

Binary Cross-Entropy without class weights, trained on the non-augmented dataset, achieves the best micro-F1 score, indicating stronger overall performance on the majority labels. In contrast, BCE with class weights trained on the non-augmented dataset yields the highest macro-F1 score, suggesting that weighted loss mitigates label imbalance and benefits minority classes. This shows a clear trade-off between overall accuracy and minority-class sensitivity.

Future work may explore augmentation methods that better preserve argumentative structure, such as supporting evidence and contradiction-based variants for each instance. Another direction is to explicitly model tweet-fallacy interactions by learning representations for both the tweet and fallacy descriptions via the [CLS] embeddings and measuring their relationships using similarity-based metrics.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o to correct typos and grammatical mistakes. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, B. Schölkopf, Logical fallacy detection, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022, pp. 7180–7198. URL: https://aclanthology.org/2022.findings-emnlp.532/. doi:10.18653/v1/2022.findings-emnlp.532.

[2] M.-H. Yeh, R. Wan, T.-H. K. Huang, Cocolofa: A dataset of news comments with common logical fallacies written by llm-assisted crowds, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 660–677. URL: https://aclanthology.org/2024.emnlp-main.39/. doi:10.18653/v1/2024.emnlp-main.39.

[3] E. Cantín Larumbe, A. Chust Vendrell, Argumentative fallacy detection in political debates, in: Proceedings of the 12th Workshop on Argument Mining, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 369–373. URL: https://aclanthology.org/2025.argmining-1.36/. doi:10.18653/v1/2025.argmining-1.36.

[4] A. Ramponi, A. Daffara, S. Tonelli, Fine-grained fallacy detection with human label variation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 762–784. URL: https://aclanthology.org/2025.naacl-long.34/. doi:10.18653/v1/2025.naacl-long.34.

[5] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[6] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), CEUR Workshop Proceedings, 2019, pp. 312–317. URL: https://aclanthology.org/2019.clicit-1.47/.

[7] J. Lanchantin, T. Wang, V. Ordonez, Y. Qi, General multi-label image classification with transformers, 2020. URL: https://arxiv.org/abs/2011.14027. arXiv:2011.14027.

[8] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, H. Wang, Sgm: Sequence generation model for multi-label classification, 2018. URL: https://arxiv.org/abs/1806.04822. arXiv:1806.04822.

[9] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news articles, 2019. URL: https://arxiv.org/abs/1910.02517. arXiv:1910.02517.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: https://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[11] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, 2019. URL: https://arxiv.org/abs/1901.05555. arXiv:1901.05555.

[12] F. Borra, C. Savelli, G. Rosso, A. Koudounas, F. Giobergia, MALTO at SemEval-2024 task 6: Leveraging synthetic data for LLM hallucination detection, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1678–1684. URL: https://aclanthology.org/2024.semeval-1.240/. doi:10.18653/v1/2024.semeval-1.240.

[13] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322–1328. doi:10.1109/IJCNN.2008.4633969.

[14] J. Li, Synthetic minority oversampling technique based on adaptive local mean vectors and improved differential evolution, IEEE Access 10 (2022) 1–1. doi:10.1109/ACCESS.2022.3187699.

[15] M. Poliakov, N. Shvai, Missynth: Improving missci logical fallacies classification with synthetic data, 2025. URL: https://arxiv.org/abs/2510.26345. arXiv:2510.26345.

[16] L. Mouchel, D. Paul, S. Cui, R. West, A. Bosselut, B. Faltings, A logical fallacy-informed framework for argument generation, 2025. URL: https://arxiv.org/abs/2408.03618. arXiv:2408.03618.

[17] S. Saha, P. Yadav, L. Bauer, M. Bansal, ExplaGraphs: An explanation graph generation task for structured commonsense reasoning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7716–7740. URL: https://aclanthology.org/2021.emnlp-main.609/. doi:10.18653/v1/2021.emnlp-main.609.

[18] A. Glazkova, O. Zakharova, Evaluating llm prompts for data augmentation in multi-label classification of ecological texts, in: 2024 Ivannikov Ispras Open Conference (ISPRAS), IEEE, 2024, p. 1–7. URL: http://dx.doi.org/10.1109/ISPRAS64596.2024.10899128. doi:10.1109/ispras64596.2024.10899128.

[19] F. Wang, M. Beladev, O. Kleinfeld, E. Frayerman, T. Shachar, E. Fainman, K. L. Assaraf, S. Mizrachi, B. Wang, Text2topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities, 2023. URL: https://arxiv.org/abs/2310.14817. arXiv:2310.14817.

[20] Ç. Çöltekin, M. Brivio, F. Can, Tübingen at politicit: Exploring SVMs, pretrained language models, and linguistic transfer for ideology detection in social media, in: M. G. Garcia, M. Passarotti (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), September 7–8, Parma, Italy, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper10.pdf. doi:urn:nbn:de:0074-3473-10. arXiv:3473/paper10.

[21] H. Fallah, P. Bellot, E. Bruno, E. Murisasco, Adapting Transformers for Multi-Label Text Classification, in: CIRCLE (Joint Conference of the Information Retrieval Communities in Europe) 2022, Samatan, France, 2022. URL: https://hal.science/hal-03727927.

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

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[24] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[25] 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, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[26] A. Ramponi, S. Tonelli, FadeIT at EVALITA 2026: Overview of the fallacy detection in italian social media texts task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[27] J. Xing, D. Luo, C. Xue, R. Xing, Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective, 2025. URL: https://arxiv.org/abs/2411.14654. arXiv:2411.14654.

[28] J. L. Ba, J. R. Kiros, G. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

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

[30] X. Zhang, Y. Song, F. Zuo, X. Wang, Towards imbalanced large scale multi-label classification with partially annotated labels, 2023. URL: https://arxiv.org/abs/2308.00166. arXiv:2308.00166.

A. Model configurations and training hyperparameters

The table below explains all model configurations that are used for experiments and official runs.

Table 3
Model configurations and training hyperparameters used in our experiments.

Component	Value
Base models	AlBERTo, UmBERTo, XLM-RoBERTa-base
Loss functions	BCEWithLogitsLoss, Weighted BCEWithLogitsLoss, Focal Loss
Optimizer	AdamW
Dropout rate	0.2
Epochs	10
Batch size	16
Learning rate	2×10^{-5}
Learning rate scheduler	Cosine schedule with linear warm-up
Warm-up steps	0.1× total training steps
Total training steps	$\lceil N_{\text{train}}/(Batchsize) \rceil \times Epochs$
Classification threshold	0.5
Focal loss α	0.25
Focal loss γ	2.0
Focal loss reduction	Mean

B. Prompt used for generating augmented data

*Sei incaricato di espandere un dataset multilabel di fallacie logiche tramite **parafrasi controllate**.*

OBIETTIVO

Dato un file JSON contenente un dataset annotato con fallacie logiche, devi generare nuove istanze sintetiche per aumentare la rappresentazione delle classi minoritarie.

PROCEDURA

1. *Conta il numero di esempi associati a ciascuna label `labels_a1` (considera esclusivamente questo set di label).*
2. *Identifica le label con meno di 100 esempi.*
3. *Per ciascuna label rara:*
   - *Considera **tutti** gli esempi reali che contengono quella label.*
   - *Per ogni esempio reale, genera **due parafrasi differenti**.*
   - *Ogni parafrasi deve essere:*
     - *estremamente diversa in struttura e lessico;*
     - *breve (1–2 frasi);*
     - *semanticamente equivalente all'originale;*
     - *coerente con la fallacia associata alla label;*
     - *naturale e plausibile come post o commento online;*
     - *una riscrittura fedele, non un'estensione, interpretazione o spiegazione.*
4. *Aggiungi ogni parafrasi al dataset.*
5. *Produci il nuovo dataset JSON completo.*
6. *Mostra il conteggio delle label prima e dopo l'augmentation.*
7. *Genera un file JSON scaricabile denominato: `dataset_augmented.json`.*

REGOLE DI GENERAZIONE

- *Nessun tono moralizzante.*
- *Nessun contenuto violento o sessuale superfluo.*
- *Nessun riferimento esplicito alla generazione o alla parafrasi.*
- *Mantieni il significato dell'input, modificando forma, struttura e vocabolario.*