MINDS at DeSegMa-IT: Detecting Human-LLM Authorship Switches via Token-Level Classification

Flavio Giobergia

Politecnico di Torino. Turin, Italy

Abstract

We address the DeSegMa-IT shared task at EVALITA, which aims to identify and segment mixed-authorship Italian text, where generation switches from a human prompt to a large language model continuation. We address the problem as a token-level sequence labeling, where each token is labelled as either human- or machine-generated. To this end, we use an encoder-only transformer, fine-tuned for token-level binary token classification (human vs. LLM), with a post-hoc threshold tuning step to improve boundary localization. Our system achieves competitive performance, ranking second in the official evaluation. In this report, we summarize the proposed solution, and explore additional options considered during the challenge.

Keywords

machine-generated text detection, authorship segmentation, large language models

1. Introduction

In recent years, rapid progress in generative artificial intelligence has led to the widespread adoption of large language models (LLMs) capable of producing text that is often difficult to distinguish from human-written content. Such models are now routinely employed in applications ranging from creative writing [1] and customer support tasks [2], to the drafting [3] and reviewing [4] of scientific papers. However, alongside these benefits, the usage of LLMs raises increasing concerns related to transparency, accountability, factual validity and misuse. Indeed, language models can be used to generate misinformation, fake news, conspiracy theories and hate content at a previously impossible rate. Such contents already flood places such as news outlets or social media. In response to these risks, regulatory frameworks – most notably the European AI Act – are increasingly emphasizing the need for mechanisms that automatically recognize machine-generated contents reliably.

Within this context, the DeSegMa-IT [5] shared task at EVALITA 2026 [6] addresses the challenge of machine-generated text detection for the Italian language, with a particular focus on "non-ideal" settings. Rather than operating under fully controlled conditions, the task is designed to evaluate the robustness of detection systems in scenarios where prior knowledge of the generation process may be unavailable. Among the proposed subtasks, Task B focuses on a segmentation problem: given a single text composed of an initial human-written prefix followed by an LLM-generated continuation, the goal is to identify the precise point at which authorship changes.

Unlike document-level detection, SubTask B explicitly emphasizes *localization*: the solutions evaluated using the mean absolute error between the predicted and gold switch *character position*. The approach of detecting an event at finer-granularities is commonly adopted in LLM-generated contents: other works in literature focused, for instance, on the detection of token-level toxic content, or the detection of hallucinations at the character level [7]. In this work, we address SubTask B by formulating it as a token-level problem, using an encoder-only transformer model (based on BERT [8]), fine-tuned on a token-level classification task. The model assigns authorship labels to individual tokens, from which the switch position is derived, allowing for fine-grained modeling of the human–machine boundary.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
flavio.giobergia@polito.it (F. Giobergia)
0000-0001-8806-7979 (F. Giobergia)

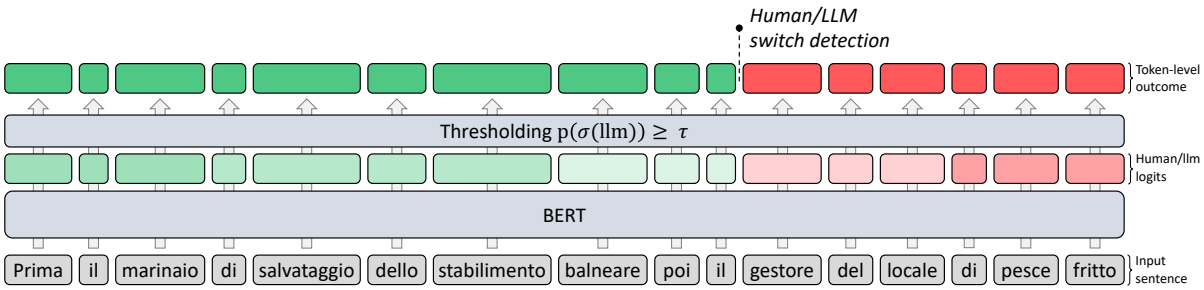

Figure 1: Intuition behind the proposed solution: tokenized sentences are annotated by a suitably fine-tuned token-level classifier and later converted to either human- or machine-generated.

2. Methodology

The proposed methodology is based on the idea of letting a model distinguish between human and LLM-generated tokens, at a token level. In other words, we build a transformer-based model capable of annotating each token of a sentence as either human- or LLM-generated. The transformer architecture is particularly well-suited for this kind of tasks, where the model can only assign a valid token-level prediction by considering the surrounding context.

We summarize the intuition in Figure 1, which shows the main steps of the proposed solution: a first token-level prediction, the following conversion to a binary outcome (human/LLM), and the final identification of the switch point (i.e., the first transition from human-predicted to machine-predicted token).

2.1. Task formulation

Let a text be tokenized into a sequence $(t_1, t_2, \ldots, t_n)$. The goal is to predict the switch point where the text transitions from human-written to LLM-generated. We frame the problem as binary token classification by assigning each token t_i a label $y_i \in \{0, 1\}$, where 0 denotes *human* and 1 denotes *LLM*. The gold labeling is derived from the annotated switch: tokens before the switch are labeled 0 and tokens at/after the switch are labeled 1.

A transformer encoder (BERT-style) produces contextual token representations $\mathbf{h}_i$, followed by a linear classification head yielding logits $\mathbf{z}_i \in \mathbb{R}^2$ and probabilities $\mathbf{p}_i = \text{softmax}(\mathbf{z}_i)$.

2.2. Switch prediction

At inference time, we obtain token-level predictions $\hat{y}_i = \mathbb{1}(\mathbf{p}_{i,1} \geq \tau)$. In other words, we assign a class based on whether the probability for LLM-generated is above a certain threshold τ (hyperparameter). We then estimate the switch token index as the earliest position predicted as LLM:

$$\hat{k} = \min\{i \mid \hat{y}_i = 1\},$$

with a fallback to $\hat{k} = n$ if no token is predicted as LLM. The final system output is the corresponding *character position* in the original text, obtained by mapping the predicted switch token back to the position of the token's starting character.

3. Experimental Results

In this section we report the main results obtained for the challenge. Specifically, we report three aspects of interest: the main results obtained during the challenge (Section 3.1), some additional experiments to identify the most suitable backbone model (Section 3.2) and, finally, a sensitivity analysis to define the hyperparameter τ (Section 3.3). We note that, for experiments that were not submitted during the competition, we report the results based on a separate test set that has been made available throughout the challenge.

Table 1
Official results for the DeSegMa-IT challenge (SubTask B). The best solution is shown in **bold**, the second best solution is underlined.

Position	Participant	MAE ↓
1	Stochastic Gradient Descenders	**52.54**
2	*MINDS (ours)*	56.53
3	Gradient Descenders	62.66
4	UniTor	81.6
-	*Regression baseline*	90.54
5	Nicla	102.04
-	*Naive prediction (avg)*	103.53

Table 2
Backbone comparison under the same token-classification framework. The best solution is shown in **bold**, the second best solution is underlined.

Architecture	Size (M params)	Pre-training language	MAE ↓
BERT	110	Italian	**56.53**
RoBERTa	125	Italian	56.95
RoBERTa	355	English	57.95
RoBERTa	125	English	58.99
BERT	340	English	76.92

3.1. Main results

Table 1 reports the official leaderboard results. Our system achieves second place overall, indicating that the proposed token-classification formulation is effective for localizing human-to-LLM transitions in Italian text. We additionally introduce (i) a simple model that addresses the task directly as a regression problem, i.e. using the output for the CLS token, after proper fine-tuning, to predict the switch point, and (ii) the performance for the naive solution that predicts the average starting point, to provide a baseline result.

3.2. Backbone definition

To understand the impact of the underlying encoder, we also compare different backbone models under the same training setup. Table 2 summarizes results across an Italian-trained BERT-style model [9] and RoBERTa-based [10] alternatives (including an Italian-adapted variant [11]). All trainings have been performed for a total of 10 epochs (batch size of 8). Further training did not provide additional generalization benefits, neither to the smaller, nor to the larger models. Overall, the results show that the Italian-tuned BERT model outperforms other models, even larger ones. This appears to indicate that, for this problem, the pre-training is more beneficial than an increased model size. Based on these conclusions, the Italian version of BERT[1] has been used as the backbone for the proposed method.

3.3. Sensitivity analysis

On top of the hyperparameters already defined for the training phase, one additional value of interest is the threshold τ applied to the model's predicted probabilities, to assign a hard label to each token. This hyperparameter can be tuned *after* the training process, using a separate validation set and studying the change in performance on the validation set, as a function of τ. We report this analysis in Figure 2. The result highlights how the model achieves good performance for small values of τ – indicating that the model is not well-calibrated. There is a limited plateau of stable performance. As the threshold

[1] `dbmdz/bert-base-italian-cased` on HuggingFace

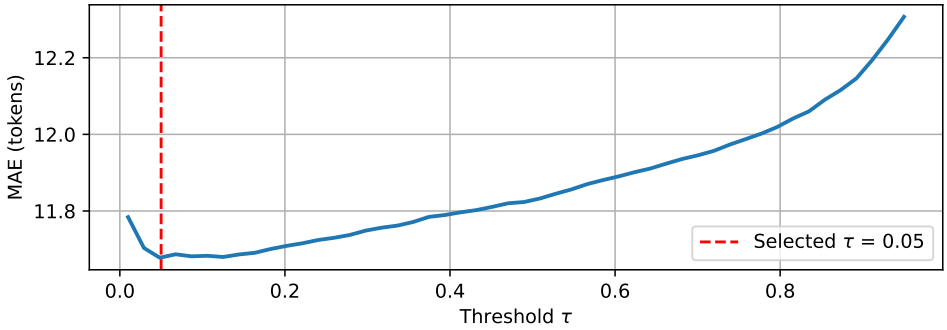

Figure 2: Performance of the model, as a function of τ. The performance is reported in terms of Mean Absolute Error computed at the token level, for consistency with the token-level approach of the proposed solution.

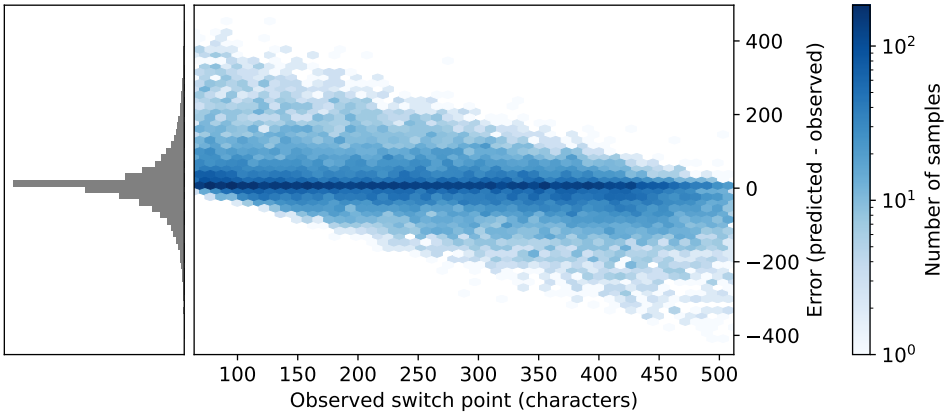

Figure 3: Representation of the residual error, as a function of the observed switch point. The side histogram shows the shape of the distribution of the errors, qualitatively representing a symmetric distribution.

increases, we observe a drop in performance (i.e., an increase in MAE): because of these empirical observations, we set the threshold to $\tau = 0.05$.

3.4. Error analysis

We additionally look into the distribution of residual errors, as shown in Figure 3. The plot shows the residual error of the model (computed as the difference between predicted and observed values), as a function of the observed switch point. The high-density region of space centered around 0 indicates that, for a majority of samples, the model behaves consistently across different switch positions. As expected, the error for "early" switch points is mostly positive (as the model can only be mistaken when predicting late switch points. The same applies for "late" switch points and reversed considerations (i.e., the mistakes are related to predicting an early switch point). We observe that the distribution of errors is symmetric: this intuitively indicates that the model does not overly predict early (or late) switches. Finally, we note that the mean (signed) error approximately 7.95: in other words, since the error is not zero-centered, the additional offset could be subtracted (e.g., if computed on an i.i.d. validation set) to compensate this bias.

4. Conclusions

We presented a token-level approach for the DeSegMa-IT task, casting authorship transition detection as binary sequence labeling with a BERT-style encoder. Our system achieves competitive performance, ranking second in the shared task evaluation. Future work could explore alternative boundary-aware objectives, improved token-to-character alignment strategies, and extensions to other languages or

mixed-generation settings. We attempted additional approaches, e.g. an alternative version of the proposed approach, with weighted importance assigned to the tokens based on their distance from the switch point. However, this alternative achieved comparable performance to the non-weighted version, and we did not report those results for simplicity. Other options, e.g. the direct prediction of the switch point as a regression task, were not found to be promising.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT 5.2) in order to: Drafting content. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] C. Gómez-Rodríguez, P. Williams, A confederacy of models: A comprehensive evaluation of llms on creative writing, arXiv preprint arXiv:2310.08433 (2023).

[2] V. Scotti, M. J. Carman, Llm support for real-time technical assistance, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 388–393.

[3] W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, et al., Mapping the increasing use of llms in scientific papers, arXiv preprint arXiv:2404.01268 (2024).

[4] F. Giobergia, A. Koudounas, E. Baralis, Large language models-aided literature reviews: A study on few-shot relevance classification, in: 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2024, pp. 1–5.

[5] G. Puccetti, A. Pedrotti, A. Esuli, Desegma-it at evalita 2026: Overview of the detection and segmentation of machine generated text in italian task, 2026.

[6] 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.

[7] C. Savelli, A. Koudounas, F. Giobergia, MALTO at SemEval-2025 task 3: Detecting hallucinations in LLMs via uncertainty quantification and larger model validation, in: S. Rosenthal, A. Rosá, D. Ghosh, M. Zampieri (Eds.), Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 1318–1324. URL: https://aclanthology.org/2025.semeval-1.175/.

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

[9] B. Staatsbibliothek, S. Schweter, bert-base-italian-cased (revision 843e404), 2025. URL: https://huggingface.co/dbmdz/bert-base-italian-cased. doi:10.57967/hf/5850.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[11] roberta-base-italian, 2025. URL: https://huggingface.co/osiria/roberta-base-italian.