Analyzing the Impact of LLM-Based Augmentation on Document Classification

Vittorio Torri[1], Francesca Ieva[1,2]

[1]MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano – {name.surname@polimi.it}
[2]HDS - Health Data Science Centre, Human Technopole

Abstract

The scarcity of annotated data remains a major limitation for supervised natural language processing, particularly in domains characterized by strong class imbalance and high annotation costs, such as the medical domain. Large Language Models (LLMs) have recently emerged as a promising tool for synthetic data augmentation, yet their effectiveness remains inconsistent across tasks and datasets. In this work, we investigate the impact of LLM-based augmentation on clinical document classification, with a focus on class-level heterogeneity. Using a dataset of Italian clinical texts, we evaluate synthetic samples generated by two open-source LLMs under different prompting strategies within a controlled TF-IDF-based classification model. Our results show that augmentation effects vary substantially across classes: while some exhibit improvements, others experience performance degradation, often reflecting a trade-off between precision and recall. To better understand these patterns, we relate class-level performance gains to measures of textual variability, including lexical diversity (MSTTR) and representation-based variance (average pairwise cosine distance) computed on TF-IDF and Sentence-BERT embeddings. We observe that, for LLaMAntino-3, augmentation gains tend to be negatively correlated with synthetic data variability, suggesting that higher lexical and semantic dispersion is associated with lower performance improvements. This trend is weaker for variance on real data and does not clearly emerge for LLaMAntino-2, which exhibits more heterogeneous behavior. Overall, our findings indicate that the effectiveness of synthetic augmentation is model-dependent and sensitive to the statistical properties of generated data, highlighting the need for class-level diagnostics and controlled generation strategies in clinical settings.

Keywords: synthetic data generation, data augmentation, large language models, document classification

1. Introduction

The availability of high-quality annotated data remains a central limitation for supervised learning in natural language processing. Although pre-trained language models have substantially reduced annotation requirements through transfer learning (Hosna et al., 2022), their performance in specialized domains or in low-frequency classes still depends critically on the presence of sufficiently representative labeled examples. This challenge is particularly pronounced in medical settings, where annotation is costly, requires domain expertise, and is often constrained by privacy regulations, leading to small and highly imbalanced datasets (Liang et al., 2017; Wu et al., 2022).

To mitigate data scarcity, a wide range of textual data augmentation strategies has been proposed. Early approaches rely on heuristic perturbations, including token-level operations such as insertion, deletion, or synonym substitution (Feng et al., 2021), as well as sentence-level transformations like back-translation (Edunov et al., 2018). While these methods can increase training diversity, their effectiveness is often task-dependent and limited by their inability to capture deeper semantic variability. Empirical evidence has shown that such techniques do not

consistently yield performance improvements and may even degrade model performance by introducing unrealistic or label-inconsistent samples (Longpre et al., 2020; Tu et al., 2020).

The recent emergence of Large Language Models (LLMs) has opened new possibilities for data augmentation. Thanks to their ability to generate fluent and contextually coherent text under zero- or few-shot prompting, LLMs have been increasingly used for both weak supervision (Wang et al., 2021) and synthetic data generation (Yoo et al., 2021). In principle, these models can compensate for missing training examples, especially in under-represented classes, by leveraging knowledge acquired during pre-training. However, despite their expressive power, LLM-based augmentation does not guarantee consistent gains. Prior studies report highly variable outcomes across tasks and datasets, suggesting that augmentation effectiveness depends on factors that are not yet fully understood (Li et al., 2023).

While prior work has evaluated augmentation at the dataset level, limited attention has been paid to class-level heterogeneity. In this study, we address this gap by investigating the role of class-level variability in determining the effectiveness of synthetic augmentation. Rather than relying on aggregate performance, we analyze how augmentation impacts individual classes within the same dataset. Our central hypothesis is that statistical properties of both real and generated text, such as lexical and semantic dispersion, influence how synthetic samples interact with the decision boundaries learned by the model.

We conduct an empirical study on a dataset of Italian clinical documents, where data scarcity and class imbalance are particularly pronounced. Our analysis reveals that augmentation effects are highly heterogeneous: while some classes benefit from synthetic data, others experience performance degradation. To better understand these discrepancies, we relate augmentation gains to multiple measures of text variance computed on both real and synthetic data.

Overall, our findings highlight that the impact of synthetic augmentation is strongly context-dependent and cannot be reliably predicted by simple heuristics. In particular, we show that the relationship between data variability and performance is non-trivial and model-dependent, emphasizing the need for class-level diagnostics when applying LLM-based augmentation. These results caution against the indiscriminate use of synthetic data and point toward more principled approaches that account for multiple properties of the generated data.

2. Methods

2.1. Classification Model

We consider a multi-class document classification problem in which the class distribution is markedly skewed, with several underrepresented categories in the training data.

To evaluate the effect of data augmentation in a controlled and interpretable setting, with a limited amount of training data, we adopt a linear classification pipeline based on TF-IDF features (Manning et al., 2008) and logistic regression. The feature space is constructed over uni-, bi-, and tri-grams, providing a balance between expressiveness and interpretability while allowing us to isolate the contribution of augmented data without introducing additional non-linear modeling effects.

Let V denote the vocabulary of all n-grams (with $n \in \{1,2,3\}$) extracted from the corpus. To reduce noise and improve robustness, we discard rare n-grams that appear in fewer than 5 documents. Each document d is represented as a vector $\mathbf{x}_d \in \mathbb{R}^{|V|}$, where each component corresponds to an n-gram $t \in V$, weighted using TF-IDF:

$$x_{d,t} = \mathrm{tf}(t,d) \cdot \log\left(\frac{N}{df(t)}\right) \tag{1}$$

where $\mathrm{tf}(t,d)$ denotes the term frequency of t in document d, $df(t)$ is the number of documents in which t appears (document frequency), and N is the total number of documents in the corpus.

Given $\mathbf{x}_i$ and class label $y_i \in \{1,\ldots,C\}$, we model class probabilities using multinomial logistic regression:

$$P(y_i = c \mid \mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}_c^\top \mathbf{x}_i)}{\sum_{c'=1}^{C} \exp(\boldsymbol{\beta}_{c'}^\top \mathbf{x}_i)} \tag{2}$$

where $\boldsymbol{\beta}_c \in \mathbb{R}^{|V|}$ are class-specific coefficient vectors. To ensure identifiability, one class is taken as reference (e.g., $c = C$) and its coefficient vector is set to $\boldsymbol{\beta}_C = \mathbf{0}$. The model is estimated by minimizing a regularized cross-entropy loss. Performance is evaluated using stratified 10-fold cross-validation, computing per-class precision, recall, and F1-score.

2.2. Synthetic Data Augmentation

To alleviate the effects of the class imbalance, we leverage generative data augmentation based on open-source large language models, specifically LLaMAntino 2 (7B) (Basile et al., 2023) and LLaMAntino 3 (8B) (Polignano et al., 2026), that are the Italian versions of the LLaMA 2 and 3 models, respectively. These models are prompted to produce synthetic documents conditioned on a target class label. We explore different prompting regimes that vary in the amount of supervision provided to the LLM. In the zero-shot setting, the model is instructed to generate a document corresponding to a given class (e.g., a clinical note for an oncology patient) without additional examples. In the few-shot settings, we augment the prompt with one or two representative documents sampled from the training set (one-shot and two-shot, respectively), thereby providing contextual guidance on style and content. This allows us to assess how varying levels of in-context examples influence both the quality and diversity of the generated samples. Generation is performed with a fixed temperature of 0.5, representing a trade-off between diversity and coherence.

For each augmented class, we generate a number of synthetic instances up to 40% of the original class size. To maintain a minimum level of data quality, we apply a lightweight post-processing pipeline that removes near-duplicate samples and filters out degenerate outputs. In particular, we discard texts containing excessive repetition (cases where a single token accounts for more than 20% of the total token occurrences) or duplicated sentence patterns, which are typical failure modes of generative models.

2.3. Assessment of augmentation effect

To quantify the benefit of augmentation, synthetic samples are progressively incorporated into the training folds, while evaluation is always performed on real (non-augmented) data. For each class, we track the change in performance relative to a baseline model trained exclusively on original data. This incremental injection protocol enables us to characterize not only whether augmentation is beneficial, but also how performance evolves as a function of the amount of synthetic data.

A central aspect of our analysis is understanding how the variability of textual data relates to augmentation effectiveness. Since there is no universally accepted measure of text diversity, we

consider three complementary metrics capturing different levels of abstraction.

First, we quantify lexical diversity using the Mean Segmental Type-Token Ratio (MSTTR) (Johnson, 1944). Given a concatenated corpus of documents, the text is partitioned into K segments of equal length $(S_1, \ldots, S_K)$. For each segment, we compute the ratio between the number of unique tokens and the total number of tokens, and average across segments:

$$MSTTR = \frac{1}{K} \sum_{k=1}^{K} \frac{|Vocab(S_k)|}{|S_k|} \tag{3}$$

This metric provides a normalized estimate of lexical richness that is less sensitive to document length than the standard type-token ratio.

Second, we assess variability in the representation space by computing the average pairwise cosine distance (APCD) (Giulianelli et al., 2020) between document vectors. We consider two complementary representations: sparse TF-IDF vectors, capturing surface-level lexical overlap, and dense embeddings obtained from Sentence-BERT (Reimers and Gurevych, 2019), which encode higher-level semantic information. For a set of N documents, the APCD is defined as:

$$Var_{APCD} = \frac{2}{N(N-1)} \sum_{i<j} \big(1 - \cos(\mathbf{v}_i, \mathbf{v}_j)\big) \tag{4}$$

where $\mathbf{v}_i$ denotes the vector representation of document i. This formulation captures the average dispersion (variance) of documents in the chosen feature space.

Finally, to investigate the relationship between data diversity and augmentation effectiveness, we compute the Spearman rank correlation coefficient between class-level performance gains and three diversity indicators: (i) the variability of the original data (Var_{real}), (ii) the variability of the generated data (Var_{syn}), and (iii) their ratio (Var_{syn}/Var_{real}). This analysis allows us to examine whether augmentation is more effective in settings and classes where synthetic data introduces additional variability, or conversely, where it better aligns with the structure of the original distribution.

Each variance measure is associated with a confidence interval obtained via bootstrap resampling. To account for uncertainty in the estimation of diversity metrics, we propagate this uncertainty to the correlation analysis. At each resampling iteration, we recompute the Spearman rank correlation with the observed performance gains. Repeating this process yields an empirical distribution of correlation coefficients, from which confidence intervals are derived using percentile estimates.

To further characterize potential differences between real and generated texts, we additionally compared the morphosyntactic structure of original and synthetic corpora through part-of-speech (POS) distributions, compute with the spaCy `it_core_news_lg` model for the Italian language (Honnibal et al., 2020). Morphosyntactic divergence was quantified using Jensen–Shannon (JS) distance computed on POS frequency distributions, as JS provides a symmetric and bounded measure for comparing discrete probability distributions.

3. Results

3.1. Dataset

For our analysis, we use as a case study the dataset proposed in (Torri and Ieva, 2026), corresponding to the Italian subset of the European Clinical Case Corpus (E3C) (Magnini et al., 2022), filtered to retain clinically relevant documents and enriched with document-level annotations, including a category label representing the medical specialty associated with the patient's condition. As described in (Torri and Ieva, 2026), these specialty labels were assigned through an LLM-based classification framework grounded on a predefined list of medical specialties and validated against a manually annotated subset of documents.

Overall, the dataset comprises 2,040 documents with heterogeneous length (median = 391 characters, IQR = 924), distributed across 14 categories with pronounced class imbalance. The most frequent categories are *Gastroenterology* (275), *Infectious Diseases* (228), and *Pediatrics* (175), while the least represented include *Hematology* (67), *Oncology* (66), and *Nephrology* (60). Augmentation experiments were conducted on 13 diagnostic classes. The residual *Other* class was excluded due to its intrinsically heterogeneous composition, which would have complicated controlled class-conditional generation and interpretation.

3.2. Data Augmentation

Figure 1 illustrates the impact of synthetic data augmentation on a subset of representative classes, selected to highlight a range of possible response patterns. Overall, the effect of augmentation is heterogeneous across classes and configurations. A recurring pattern is a trade-off between precision gains and recall degradation, particularly evident in *Rheumatology* and *Orthopedics*, where substantial precision improvements, especially for LLaMAntino-3, are often offset by drops in recall, leading to mixed effects on F1. In contrast, *Pediatrics* shows declines in precision, with a small improvement in recall. *Nephrology* exhibits more moderate and inconsistent changes, with small gains or losses depending on the configuration. This variability is even more marked in *Pulmonology*.

Notably, the two model families exhibit systematically different behaviors: LLaMAntino-3 tends to better exploit synthetic augmentation, often achieving larger precision gains and more favorable F1-score trade-offs, whereas LLaMAntino-2 shows more frequent performance degradation. The different learning configurations do not exhibit a clear ranking, although zero-shot settings often yield both the best and the worst outcomes.

To investigate the drivers of these heterogeneous effects, Figure 2 examines the relationship between class-level variance metrics and augmentation gains (ΔF1), considering both real and synthetic data distributions.

A key observation is that LLaMAntino-3 exhibits a more consistent and coherent pattern, with correlations that are predominantly negative across variance metrics, particularly when considering synthetic variance (Panel b). Importantly, while confidence intervals remain relatively wide, the direction of the effect is stable across configurations, suggesting a consistent negative association. These results suggest that, for LLaMAntino-3, augmentation tends to be more effective when synthetic samples remain close to the original class distribution, and that excessive diversity, both lexical and semantic, can be detrimental. In contrast, LLaMAntino-2 shows more heterogeneous and less stable behavior.

When focusing on real data variance (Panel a), correlations are generally weaker and less structured. This suggests that intrinsic class variability alone is not a strong predictor of augmenta-

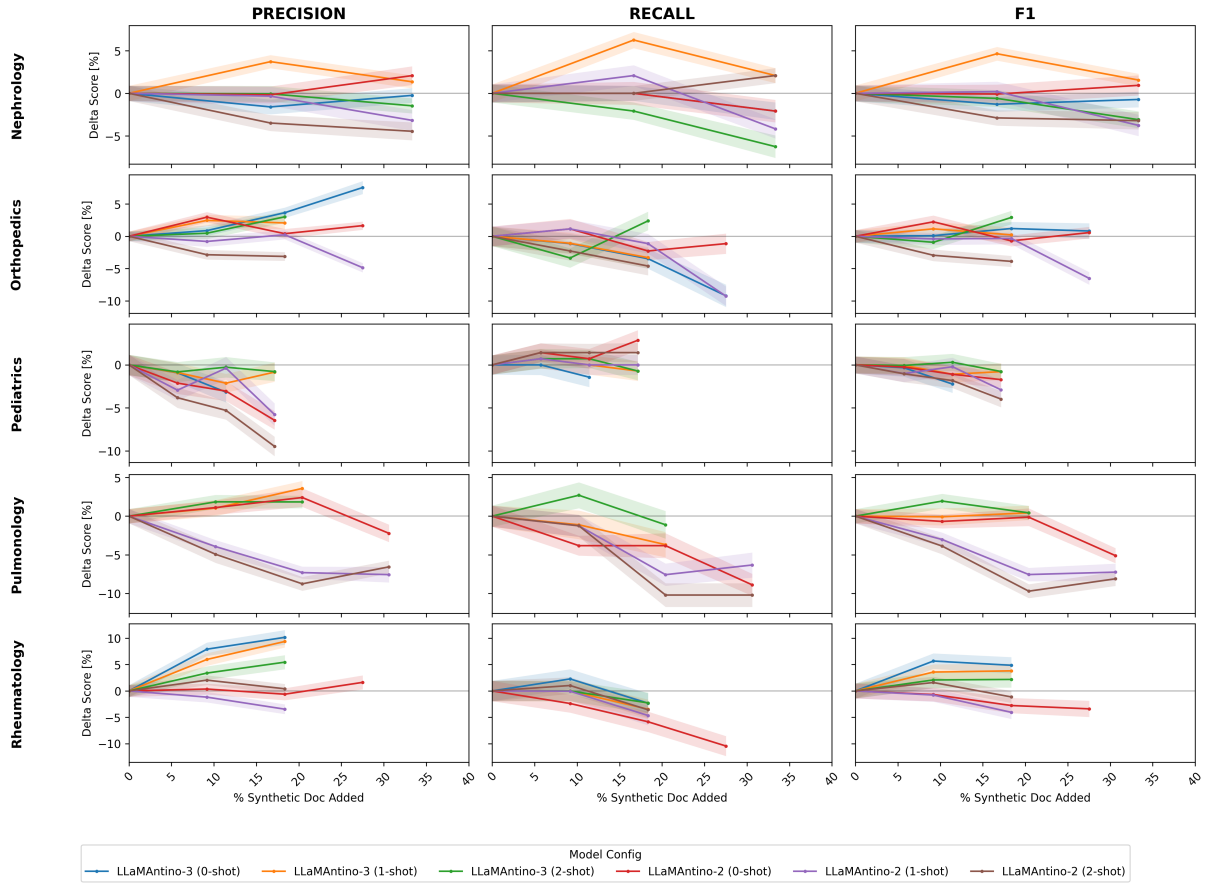

Figure 1: Effect of augmentation with synthetic documents generated by two different LLMs (LLaMAntino-2 and LLaMAntino-3) in three different settings (0-shot, 1-shot, 2-shot), on some classes of the dataset, in terms of precision, recall and F1-score (shaded areas represent 95% C.I. from 10-fold CV), varying the amount of synthetic samples added (% to original class size).

tion effectiveness, and that the properties of the generated data distribution play a more critical role.

Overall, these results highlight that the impact of variability on augmentation is model-dependent, with LLaMAntino-3 following a more interpretable pattern and a consistent relationship between generated variability and performance, indicating a diversity–fidelity trade-off, whereas LLaMAntino-2 exhibits less predictable behavior.

To further investigate whether the observed variability reflected meaningful diversity or generation artifacts, we compared the morphosyntactic structure of real and synthetic documents through part-of-speech (POS) distributions. As shown in Table 1, all generation settings exhibited only limited divergence from the original corpus, with Jensen–Shannon distances ranging from 0.0045 to 0.0095. Interestingly, few-shot prompting generally reduced morphosyntactic divergence compared to zero-shot generation. Overall, these results suggest that the generated documents broadly preserve the linguistic structure of the original clinical texts, and that the variability captured by lexical and embedding-based measures is unlikely to be explained by major morphosyntactic drift alone, suggesting that more complex semantic factors contribute to the observed augmentation effects.

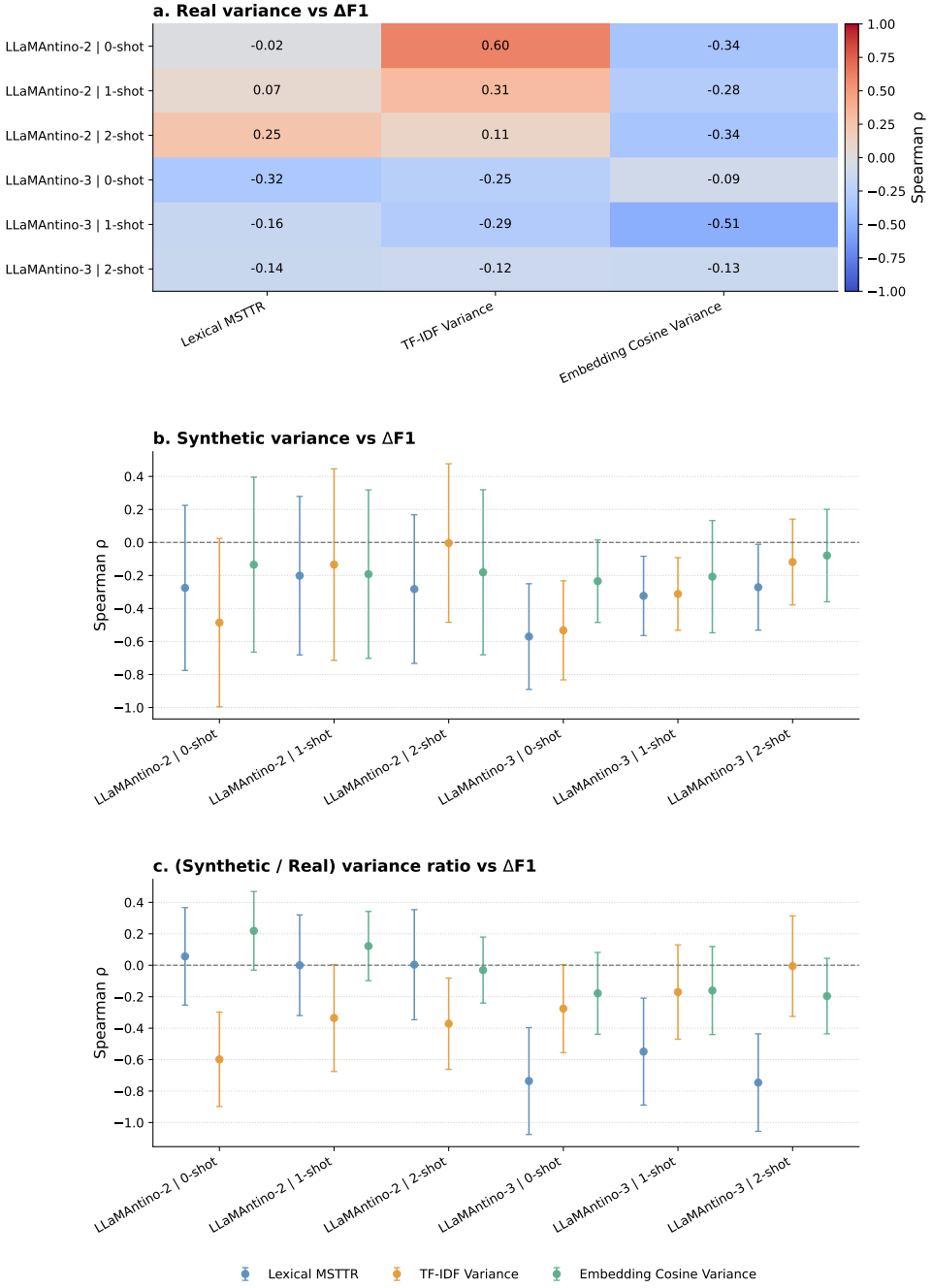

Figure 2: Relationship between class-level variance and augmentation gains (ΔF1). (a) Spearman correlation (ρ) between ΔF1 and variance computed on real data across three metrics: lexical diversity (MSTTR), TF-IDF variance, and embedding cosine variance. (b) Corresponding correlations using variance computed on synthetic data. (c) Correlation between ΔF1 and the ratio of synthetic to real variance. Error bars in (b) and (c) represent bootstrap confidence intervals.

Table 1: Jensen–Shannon (JS) distance between the part-of-speech (POS) distributions of real and synthetic corpora under different generation settings.

Model	Shots	# gen. tokens	JS Distance POS
LLaMAntino-2	0	292,518	0.0062
	1	291,180	0.0045
	2	290,422	0.0050
LLaMAntino-3	0	296,480	0.0095
	1	293,517	0.0058
	2	294,538	0.0065

4. Conclusions

In this work, we analyzed the effectiveness of LLM-based data augmentation in clinical document classification, focusing on class-level heterogeneity. Our results show that augmentation effects are highly variable: while some classes benefit, others experience performance degradation, often due to precision–recall trade-offs. This highlights the limitations of aggregate evaluation and the need for class-specific analysis.

Our analysis further suggests a potential relationship between augmentation effectiveness and the variability of synthetic data. In particular, for LLaMAntino-3, we observe a tendency toward an inverse association between performance gains and synthetic variability, suggesting that augmentation may be more effective when generated samples remain closer to the original data distribution rather than maximizing diversity. Additional morphosyntactic analyses based on POS distributions indicate that the generated texts remain broadly aligned with the linguistic structure of the original corpus, implying that the observed variability effects are unlikely to be explained by major morphosyntactic drift alone. Nevertheless, the relationship between variability and downstream performance should be interpreted cautiously, as the estimated correlations remain associated with relatively wide confidence intervals and are less consistently observed for other generative models such as LLaMAntino-2, which exhibits more heterogeneous behavior across configurations.

This study has several limitations. First, the analysis was conducted on a single Italian clinical dataset characterized by specific linguistic, stylistic, and class-imbalance properties. Consequently, the observed relationships between synthetic data variability and augmentation effectiveness may not directly transfer to other domains, languages, or document collections. Second, we adopted a relatively simple and controlled TF-IDF + logistic regression classification framework to isolate the contribution of synthetic augmentation and facilitate interpretability. In this setting, synthetic samples primarily modify class distributions and local lexical statistics within a fixed feature space, allowing a more direct analysis of the relationship between generated variability and downstream performance. However, modern embedding-based and transformer-based classifiers jointly learn representations and decision boundaries during training, potentially leading to substantially different interactions between synthetic data and model behavior. In such architectures, synthetic variability may not only introduce noise or distributional mismatch, but also reshape the geometry of the learned representation space, potentially acting as a form of regularization or improving class separation. As a consequence, the diversity–fidelity trade-offs observed in our experiments may manifest differently in those settings. Finally, we considered generative models of relatively limited size (up to 8 billion parameters), and different behaviors may emerge with larger LLMs. Future work should therefore extend

the analysis across datasets, languages, classifier architectures, and model families, while also more systematically investigating the role of generation hyperparameters (e.g., temperature) in controlling the diversity–fidelity trade-off.

Overall, our findings indicate that synthetic augmentation is not universally beneficial, and that its effectiveness depends on both the properties of the generated data and the underlying model. These results motivate more controlled and diagnostic approaches to LLM-based augmentation in data-scarce settings.

Acknowledgment

The present research has been partially supported by the Italian Ministry of University and Research (MUR), grant Dipartimento di Eccellenza 2023-2027, awarded to the Department of Mathematics, Politecnico di Milano.

References

Basile P., Musacchio E., Polignano M., Siciliani L., Fiameni G., and Semeraro G. (2023). LLaMAntino: LLaMA 2 models for effective text generation in italian language.

Edunov S., Ott M., Auli M., and Grangier D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500.

Feng S. Y., Gangal V., Wei J., Chandar S., Vosoughi S., Mitamura T., and Hovy E. (2021). A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988.

Giulianelli M., Del Tredici M., and Fernández R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3960–3973.

Honnibal M., Montani I., Van Landeghem S., and Boyd A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hosna A., Merry E., Gyalmo J., Alom Z., Aung Z., and Azim M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data*, 9(1):102.

Johnson W. (1944). Studies in language behavior: a program of research. *Psychological Monographs*, 56(2):1–15.

Li Z., Zhu H., Lu Z., and Yin M. (2023). Synthetic data generation with large language models for text classification: potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10443–10461.

Liang J. J., Tsou C.-H., and Devarakonda M. V. (2017). Ground truth creation for complex clinical NLP tasks–an iterative vetting approach and lessons learned. *AMIA Summits on Translational Science Proceedings*, 2017:203.

Longpre S., Wang Y., and DuBois C. (2020). How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4401–4411.

Magnini B., Altuna B., Lavelli A., Minard A.-L., Speranza M., and Zanoli R. (2022). European clinical case corpus. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pp. 283–288. Springer.

Manning C., Raghavan P., and Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Polignano M., Basile P., and Semeraro G. (2026). Advanced natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA. *Scientific Reports*.

Reimers N. and Gurevych I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992.

Torri V. and Ieva F. (2026). Leveraging LLMs for medical document clustering: towards a benchmark dataset. In F. D. B., Leorato S., Masci C., and Nicolussi F., editors, *Statistical Methods for Data Analysis and Decision Sciences*, Italian Statistical Society Series on Advances in Statistics. Springer. To appear.

Tu L., Lalwani G., Gella S., and He H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Wang S., Liu Y., Xu Y., Zhu C., and Zeng M. (2021). Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4195–4205.

Wu H., Wang M., Wu J., Francis F., Chang Y.-H., Shavick A., Dong H., Poon M. T., Fitzpatrick N., Levine A. P., et al. (2022). A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *npj Digital Medicine*, 5(1):186.

Yoo K. M., Park D., Kang J., Lee S.-W., and Park W. (2021). GPT3Mix: leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2225–2239.