Literature Review on Generating Synthetic Data for Text Classification

Using LLMs for Generating Synthetic Data for Text Classification

Spain

1. Introduction

Codification of text into the code of a statistical classification is a very resource-consuming task that must be done routinely in statistical offices, as well as in other governmental organizations. It is thus indispensable to devise automatic or assisted methods of codification to reduce the burden of the task. In the last decade, the fields of Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP) have seen tremendous advances, which have made possible the use of these techniques to obtain reliable codifiers which can be used profitably by statistical organizations. However, this possibility is just beginning to be explored by statistical offices, and there are some important challenges that must be overcome in order to achieve a good solution.

One of the challenges of machine learning models for text classification is the great amount of good-quality data needed to train a reliable model. This is a particularly acute problem in the context of statistical classifications, due to several factors like the complexity of the classifications, the scarcity of experts who can reliably annotate text, the possible ambiguity of the text, and the huge imbalance between classes.

Several techniques have been developed in NLP to deal with this training data scarcity problem, essentially by obtaining new synthetic samples, either by modifying available real-world samples (by using synonyms, introducing spelling errors, etc.) or by leveraging pre-trained large language models (LLMs) to generate new synthetic samples.

In this literature review, we survey some of the most relevant papers published in the last years regarding these issues, focusing especially on the use of LLMs to generate synthetic datasets. This is a very active area of research, as can be seen by the volume of the literature in the last few years, and we expect it to be a rapidly moving field since the advances in LLM technology can have a huge impact in their synthetic generation capabilities.

2. Data augmentation

Data augmentation is a technique in machine learning that faces the challenge of training models with a limited amount of data. One significant advantage of data augmentation is the ability to palliate the lack of high quality and real-world data. There are several techniques for text classification like synonym replacement, random insertion, or even more sophisticated methods which involve using LLMs, that prevent the underrepresentation or misrepresentation of classes and generate anonymous data.

Furthermore, for models trained on a bag-of-words representation, data augmentation can help train more robust and accurate classification models. In the field of free-text classification, augmenting data with slight variations in wording or sentence structure can make the model more resistant to different ways of expressing the same idea, misspellings, and previously unseen data.

However, there are also inconveniences associated with the use of data augmentation. One potential drawback is the risk of introducing noisy or misleading data if the augmentation techniques are not carefully designed. Data augmentation could change text semantic meaning and potentially mislabel the augmented sample and also could alter the distribution of the training data which may then differ from the real-world distribution.

Another challenge lies in determining the most effective augmentation techniques for a specific task and dataset. The optimal augmentation strategies can vary significantly depending on the data modality, the nature of the classification task, and the underlying model architecture. Additionally, the evaluation of augmented data quality remains a challenge, as there are no standardized metrics to directly assess the diversity and faithfulness of generated samples.

Moreover, while LLMs data augmentation enhance the generation of sophisticated and diverse content, it can also suffer from limitations such as sensitivity to the underlying models (e.g., misunderstanding prompts) and potential issues in handling complex relationships within the data.

2.1. Classical Text Augmentation Techniques

Classical text augmentation methods are mainly based on rule-based or statistical text modifications. Synonym replacement, word embeddings, and back-translation are techniques that have shown potential in train set augmentation but have limitations. For instance, synonym replacement using WordNet can improve text classification accuracy, though results vary depending on dataset and model architecture, and can be an over effort for some languages. Some studies showed that synonym replacement improved accuracy by +0.8% but others showed that it can negatively affect results [Zhou et al., 2024]. Embedding replacement methods, such as using Word2Vec or GloVe, enhance model robustness but usually introduce noise, reducing performance in classification tasks [Wang et al., 2024]. Back-translation, translating text into another language and then back to the original, helps to diversify training data while maintaining the context of the sentence. [Wang et al., 2024]

2.2 Text Augmentation with Large Language Models (LLMs)

Recent advancements in LLMs have introduced more sophisticated text augmentation techniques, such as masked language modeling (MLM) and paraphrase generation. Methods like c-BERT (contextualized BERT) improve text classification accuracy [Zhou et al., 2024]. LLMs can generate diverse and contextually rich text, making them particularly useful for low-resource languages or tasks with limited training data. Another promising approach is contrastive data augmentation, where models generate semantically similar yet distinct sentences, improving generalization in tasks like natural language inference [Bayer et al., 2022].

A hybrid approach combining classical augmentation (to introduce diversity) and LLM-based augmentation (to ensure coherence) often yields the best results [Zhou et al., 2024][Bayer et al., 2022].

3. Dataset generation-based learning

The appearance of large language models pre-trained with a massive amount of text has opened the door to new strategies for text classification. Since these models have a very good understanding of language, it is now possible to use them to directly classify text via prompting, either as a zero-shot (without real-world samples) or few-show (providing some real-world samples in the prompt) problem. This strategy can be improved significantly by providing to the LLM relevant context for the task, such as the relevant class titles, class descriptions and examples, as provided for instance in the explanatory notes of the statistical classification. This can be achieved using the technique known as Retrieval Augmented Generation (RAG) and is explored by another cluster of this project.

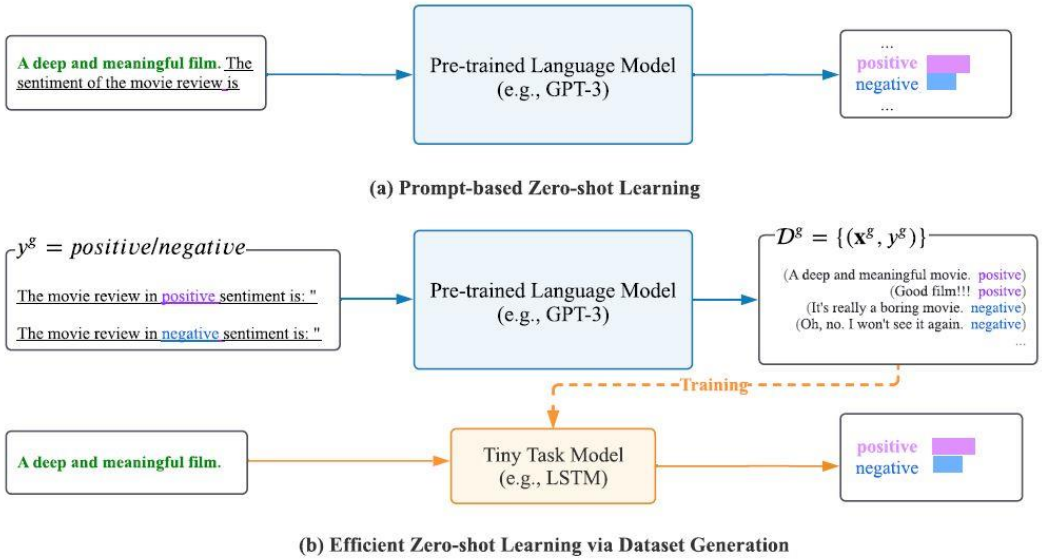
However, the direct use of LLMs for text classification in the statistical context has some drawbacks. First, there is the issue of the huge computational resources that the statistical offices need to have to be able to use one of these models in-house. As an alternative to in-house models, the models in external servers provided by private companies can be called via API. However, this approach raises issues regarding the confidentiality of the data and the monetary cost of the API use. In addition, stability of the models is another concern of this approach, since the inherently stochastic nature of the generative language models makes it difficult to achieve reproducible results.

An alternative approach to classification, which as we will see has been also extensively explored in the literature, is to leverage the language knowledge of an LLM to generate a synthetic dataset of samples (pairs text – label), and then use this synthetic dataset, maybe incremented by real-worlds samples, to train or fine-tune a smaller classification-specific model, like for instance a BERT model. This approach avoids the mentioned drawbacks of using an LLM directly as a classifier, while achieving good performance in the classification task. This is known as dataset generation-based learning.

In this section we survey some papers dealing with several variants of this last strategy.

3.1 ZEROGEN: Efficient Zero-shot Learning via Dataset Generation

The paper [Ye et al., 2022] presents ZeroGen, a straight-forward strategy similar to the one already explored by Statistics Spain in the context of NACE classification: given a zero-shot task, a first dataset is generated from scratch using large pre-trained language models (PLMs) in an unsupervised manner. Then, a tiny task model (TAM), such as LSTM, is trained under the supervision of the synthesized dataset.



This approach allows highly efficient inference as the final task model only has orders of magnitude fewer parameters compared to PLMs (e.g., GPT2-XL). This approach is annotation free (no human annotation is involved in the process) and efficient (~0.4% number of parameters).

ZeroGen can be seen as a variant of knowledge distillation. However, it does not require human annotation during the distillation process, and it makes no assumption on the architecture choice of the student model.

Three strategies are compared in this paper: ZeroGen, zero-shot prompting and supervised learning. ZeroGen surpasses zero-shot prompting on almost all datasets and even surpasses supervised learning for a small number of datasets, those with few samples.

The importance of prompt engineering is highlighted in the paper, that shows how results heavily rely on the prompting strategy and that the best approach varies from one PLM to another.

The authors also show that the size of the synthetic dataset is key, reaching a plateau of performance for three different tasks (classification, natural language inference and question answering) at around 10^5 – 10^6 samples.

As is apparent by reading many other papers, it seems that ZeroShot is considered a de facto standard and there exists a big consensus that it represents, at least, a respectable baseline for the zero-shot approaches.

3.2 Generating Training Data with Language Models: Towards Zero-Shot Language Understanding

The paper [Meng et al., 2022] presents SuperGen (Supervision Generation), a similar approach to ZeroGen. A unidirectional PLM generates class-conditioned texts guided by prompts, which are used as the training data for fine-tuning a bidirectional PLM to solve a Natural Language Understanding (NLU) task. With quality training data selected based on the generation probability and regularization techniques (label smoothing and temporal ensembling) applied to the fine-tuning stage for better generalization and stability, the

approach outperforms zero-shot prompting methods and achieves even comparable results to few-shot approaches.

One difference between SuperGen and ZeroGen is that, in SuperGen, the classification task is solved using a quite big PLM, of similar size than the one used for text generation, while ZeroGen uses a TAM.

The paper presents some techniques to improve the quality of the generated dataset:

- Selecting quality training data: the authors propose a technique to select generated samples with the highest probabilities of pertaining to the desired label. They estimate these probabilities via the generation probability conditioned on the prompt.
- Regularization for Better Generalization and Stability:
    - Label smoothing: the label is a weighted average of the one-hot vector and a uniform distribution over all labels. This approach can be improved by taking advantage of the hierarchical nature of statistical classifications, giving more weight to "closer" categories.
    - Temporal ensembling: consists of recording the predictions of the classifier on each training sample at different training steps and using the accumulated moving-average predictions to regularize the latest model training.

Finally, the paper presents some ethical considerations. PLMs can come with potential risks or harms such as generating misinformation or amplifying harmful biases. This must be taken into account when using ZeroGen, SuperGen or similar approaches in official statistics, but we think that the risks can be reduced when using official descriptions, notes and/or titles for the prompts.

3.3 PROGEN: Progressive Zero-shot Dataset Generation via In-context Feedback

The paper [Ye et al., 2022] presents a progressive zero-shot dataset generation framework, ProGen, that represents a sophistication of ZeroGen and SuperGen. It aims to improve the quality of dataset synthesis. The motivation for this paper is the observation that synthetic datasets have long been suffering from low-quality issues (e.g., low informativeness and redundancy). This explains why a massive synthetic dataset does not lead to better performance.

The authors use ZeroGen to build the backbone of their framework. Concretely, they first train a task-specific model (TAM) with a partially generated dataset. Then, assuming no access to human annotations, they estimate the influence of each sample via the noise-robust influence function. Finally, with those identified most influential samples, they explore the use of in-context learning to shift the generation distribution towards that of influential samples, so that the system generates more related samples. The whole framework progressively constructs the synthetic dataset and enhances the performance of the final task-specific model.

Across multiple text classification datasets, the authors compare various approaches. Summing up the results:

Prompting (zero-shot classification) << ZeroGen (DistilBERT) < ProGen (DistilBERT) < Supervised (DistilBERT).

They also show that any of the methods achieves much worse performance when replacing DistilBERT with an LSTM.

3.4 Self-guided noise-free data generation for efficient zero-shot learning

Experimental observations across many downstream tasks indicate that, in ZeroGen, the model trained on the PLM-generated dataset suffers from overfitting to noisy data. The authors identify two major cases of noisy samples in the synthetic dataset: corrupted labels and task-irrelevant samples.

The paper [Gao et al., 2023] proposes SunGen, an approach to enhance the quality of generated data. The authors resort to a family of noise-robust loss functions. These functions theoretically show a noise-tolerant property. However, from the optimization point of view, such loss functions suffer from instability and difficulty when training the neural networks. SunGen leverages the noise-tolerant property of these losses, while avoiding their pathology. The authors propose a novel bi-level re-weighting framework: in the inner loop, they train the task model using weighted training loss based on current sample weights; in the outer loop, the noise-robust loss is adopted to guide the learning of the sample weights. The two procedures are performed alternatively to generate a set of weights indicating the importance of samples.

The results show significant improvements over ZeroGen, like ProGen.

3.5 FuseGen: PLM Fusion for Data-generation based Zero-shot Learning

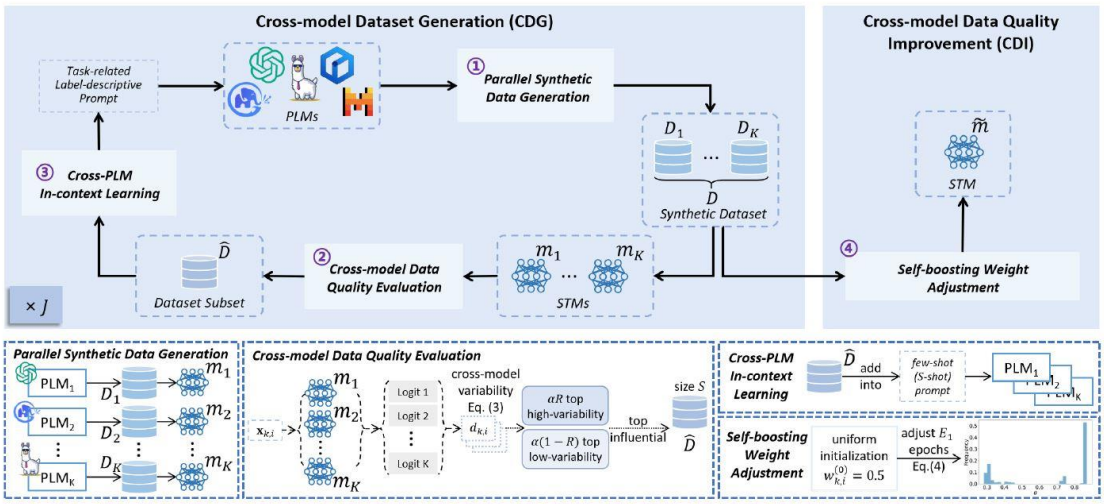
In [Zou et al., 2024], the authors present FuseGen, a strategy to improve the quality of the generated dataset. The main difference with ProGen and SunGen is that FuseGen uses more than one PLM in parallel.

In a nutshell, FuseGen consists of two main components: Cross-model Dataset Generation (CDG) and Cross-model Data Quality Improvement (CDI).

For CDG, given a fixed number of samples to generate in total, PLMs progressively generate datasets for multiple rounds, each round using an improved subset of samples generated from previous rounds as in-context examples. This is realized in three steps:

1. Parallel Synthetic Data Generation: each PLM generates its own dataset and trains a respective Small Task-specific Model (STM).
2. Cross-model Data Quality Evaluation: the quality of generated samples is evaluated using a cross-PLM criteria to select a desirable subset.
3. Cross-PLM In-context Learning: the cross-PLM subsets are used as in-context examples to prompt PLMs to generate new datasets.
4. Step (1) is then repeated.

After the required number of samples is reached, CDI is performed to re-weight samples with a self-boosting strategy.



The paper also presents a comparison between ZeroGen, ProGen, SunGen and FuseGen. Summing up the results, in average:

ZeroGen << ProGen < SunGen << FuseGen.

It seems, as expected, that the performance improves as the complexity of the approach increases.

4. Other aspects of LLM-generated synthetic data

In this section, we comment on some papers dealing with aspects of the use of synthetic data generated by LLMs. In our opinion, the first paper we discuss is the most relevant one, since it provides a generic workflow for synthetic data generation using LLMs, also giving advice on the best practices and ways to construct the prompts for generating the dataset, as well as strategies for evaluating the quality of the resulting dataset. Other papers reviewed in this section include alternative approaches to the use of LLMs for constructing synthetic datasets, strategies for generating datasets with greater diversity, use of LLMs as data annotators, and quality evaluation of synthetic datasets.

4.1 A generic workflow for synthetic data generation using LLMs

The paper [Long et al., 2024] is a recent survey which represents a first attempt to describe a unified framework and generic workflow for synthetic data generation using LLMs. The basic setting is as follows. We want to generate synthetic data $\mathbf{D}_{\mathbf{gen}}$ for a given task **T** (text classification in our case), using an LLM and a small number of real samples $\mathbf{D}_{\mathbf{sup}}$. We impose two conditions on the generated dataset $\mathbf{D}_{\mathbf{gen}}$: faithfulness (samples have relevant content and correct labels) and diversity (the generated samples capture the variation among real-world data in style, text length, etc.).

In their paper, they define a generic workflow in three stages: data generation, data curation, and data evaluation.

In the stage of data generation, a well-designed prompt must be used to generate a good-quality dataset. Such a prompt consists of three parts: task specification, which can be a prologue such as “Suppose you are an annotator for <problem>”, generation conditions,

which is a variable part of the prompt designed to provide diversity in the samples by communicating to the LLM the specific type of data desired, and in-context demonstration, which are demonstrations of the desired data, usually obtained from real-world samples $\mathbf{D}_{\mathbf{sup}}$. Moreover, more complex strategies of data generation are discussed, such as multi-step generation in which the overall generation process is manually decomposed into a chain of simpler subtasks.

The dataset generated in the previous step may have a considerable portion of noisy or worthless samples. In the stage of data curation, a high-quality subset $\mathbf{D}_{\mathbf{cur}}$ of the generated data $\mathbf{D}_{\mathbf{gen}}$ is obtained. To obtain this high-quality dataset some techniques have been used, such as high-quality sample filtering, in which some heuristic metric is used to order the generated samples according to their quality and then filter them, or label enhancement, in which either a human or an auxiliary model rectify the labels of the incorrectly labeled samples.

In the last stage, data evaluation, the quality of the synthetic dataset is evaluated to ensure its value to downstream tasks. The methods covered in the paper are divided into two categories: direct and indirect evaluation. In direct evaluation, the quality of individual samples is assessed according to their faithfulness (either by human inspection or using some auxiliary model) and to their diversity (for instance by using vocabulary statistics and sample relevance calculations). In indirect evaluation, the quality of the dataset is evaluated by its performance in downstream tasks.

4.2 LLM generation with human supervision

In [Wang et al., 2021] the authors propose the use of LLMs (in this case GPT-3) for data annotation, thus saving considerable time, cost and human resources. They explore the usefulness of GPT-3 in annotating unlabeled text samples to obtain datasets with which train smaller BERT-like models for classification. Two strategies for combining GPT-3 and human annotators are proposed: one where every sample is allocated either to a human annotator or to a GPT-3 annotator, taking into account the cost and the budget, and another one where human annotators re-annotate the data labeled with GPT-3 with the lowest confidence score.

Empirical analyses of the proposed strategies in different NLP tasks (including text classification) show that training the classifier just with only-GPT-3 annotated data surpasses the performance of the classifier trained with only-human annotated data in a low-budget regime, while a mix of GPT-3 and human annotated data increase the performance even more. Moreover, performance of small models with GPT-3 annotated data is seen to surpass that of using directly GPT-3 in a zero-shot or few-shot setting for some tasks.

In [Chung et al., 2023] the authors investigate strategies to enhance the diversity of text data generated by Large Language Models (LLMs) while preserving accuracy. The authors explore two diversification techniques: logit suppression, which reduces the likelihood of generating frequently produced language patterns, and temperature sampling, which broadens token selection probabilities. Their findings indicate that while these methods increase diversity, they often compromise the accuracy of the generated data, leading to misaligned text and labels.

To address this trade-off, the study examines two human intervention approaches: label replacement (LR), which corrects misaligned labels, and out-of-scope filtering (OOSF), which removes instances irrelevant to the target domain. The authors show that LR significantly enhances model performance, and in some cases, models trained with LR interventions outperform those utilizing LLM-based few-shot classification. Conversely, OOSF does not show a notable improvement in accuracy, suggesting the need for further research into effective human-in-the-loop strategies for text data generation.

4.3 Improving diversity by leveraging existing datasets

In [Gandhi et al., 2024], a pipeline is introduced for transforming existing datasets for a different task into a synthetic dataset for the task of interest. The basic idea is, given a task, such as classification of economic activities, to retrieve the most relevant dataset from a repository of datasets (such as Hugging Face Hub) and use an LLM to transform those datasets into a dataset tailored to our task which can then be used to fine-tune some task-specific model. The authors note that this approach leads to more diverse and difficult samples than pure synthetic generation, thus improving the performance of the classifiers. Moreover, they also note that the data generated with this method can be combined with purely synthetic data in order to boost the performance of the models.

In [Divekar et al., 2024], a novel approach to synthetic dataset generation is introduced that improves both diversity and realism in text classification tasks. Standard few-shot prompting with large language models (LLMs) often leads to repetitive outputs and biases, limiting the effectiveness of synthetic data. The method proposed by the authors addresses these issues by integrating retrieval augmentation, where external documents are used as prompts to guide LLM-generated text. This method ensures that the generated data is more lexically and semantically diverse while maintaining contextual coherence with human-authored text.

Empirical evaluations across multiple classification tasks show that this method significantly outperforms traditional prompting techniques. The paper highlights that datasets generated through retrieval augmentation lead to better downstream model performance.

4.4 Evaluation and quality of LLM-generated synthetic datasets

In [Li et al., 2023], the authors discuss some aspects regarding the quality of synthetic datasets generated by LLMs and their performance for classification tasks, by looking at the factors impacting the effectiveness of LLM-generated synthetic data in facilitating successful model training. They make an empirical evaluation of two data generation settings: zero-shot generation, in which no real-world samples are provided, and few-shot generation, in which some real-world samples are provided to the LLM (GPT 3.5-Turbo). They compare the performance of classification models (BERT and RoBERTa-based) fine-tuned with these synthetic datasets to those fine-tuned with real-world data. The conclusion is that, while models trained with high-quality real- world data surpass the performance of models trained only with synthetic data, using real-world samples in the generation process (few-shot generation) clearly surpass zero-shot generation. Moreover, it is observed that an important factor that affects performance of models trained with synthetic data is the subjectivity of the task, measured as the proportion of samples where different human annotators disagree on their label. It is noted that

synthetic data generation works significantly better in tasks with low subjectivity level, due to the incapacity of LLMs to generate enough diversity in the samples for highly subjective tasks.

The paper [Liu et al., 2024] provides a comprehensive overview of synthetic data's role in AI development, addressing its applications, challenges, and future directions. It highlights how synthetic data can mitigate issues related to data scarcity, privacy concerns, and high acquisition costs by generating artificial datasets that replicate real-world patterns. The authors present empirical evidence demonstrating the effectiveness of synthetic data across various domains, emphasizing the importance of ensuring its factual accuracy, fidelity, and impartiality. They advocate for the responsible use of synthetic data to build more powerful, inclusive, and trustworthy language models.

However, the paper also acknowledges significant challenges associated with synthetic data, such as the risk of proliferating misinformation, introducing ambiguity in AI alignment, and complicating evaluation processes due to potential contamination. To address these issues, the authors recommend establishing clear guidelines and best practices for ethical data generation, developing robust methods for validating and testing AI models trained on synthetic data, and creating proprietary evaluation benchmarks to prevent data leakage. They also suggest future research directions, including scaling synthetic data generation, enhancing its quality and diversity, and exploring the potential for AI systems to self-improve through synthetic data.

5. Synthetic data generation via LLMs in official statistics

As far as the authors are aware, at the time of writing the only explorations of the ideas of this document in the context of official statistics are the work of the Statistical Unit of the german Federal Employement Agency [Heß, 2024], the work of the Department of Statistics of Singapore [Lim et al.,, 2025] and the work of Statistics Spain [Pérez-Bote et al., 2025] in economic activity classification (NACE). Therefore, there is plenty of room for further exploration and improvement of these ideas. In this section we discuss some of the peculiarities of the text classification problem in the context of official statistical classifications, and the work done by Statistics Spain.

The problem of text classification in the context of statistical classifications is a highly complex one and differs substantially in several aspects from other problems treated in NLP. Statistical classifications are usually very complex classifications with hundreds of classes and a rich hierarchical structure, and some of the classes are very scarcely populated. Moreover, the boundaries between classes are not always clear to the non-expert, making it very difficult to correctly classify some of the samples.

The field-collected samples are typically not of good quality, being usually short descriptions, sometimes with grammatical mistakes, and even in some cases not providing enough information for being reliably classified in one of the classes of the statistical classification. Moreover, since some of these samples are not codified by classifications experts, there is a non-negligible proportion of misclassified samples. Another problem is that real-world data is highly imbalanced, with some classes having very few real-world samples.

Another important distinction with other NLP problems is the availability of very high-quality information on statistical classifications, such as that provided by classes titles and explanatory notes, which typically provide some examples of each class and also some non-examples which a non-expert could incorrectly classify in a given class. This is a valuable resource for model training, which is not available in many other NLP problems.

By the reasons just explained, the issues of real-world data have a negative impact in models trained exclusively using this data. Data augmentation techniques, and particularly synthetic generation using LLMs, can provide more reliable data, especially if using the high-quality information provided by explanatory notes and other statistical documentation.

In [Heß, 2024], several models are developed for the German version of the NACE classification using the dataset generation-based learning approach from section 3. First, descriptions of economic activities from the explanatory notes are tokenized. GPT-4 Turbo is then used to generate synthetic samples of economic activities, some of them based on the text from the explanatory notes, and some of them based on a list of keywords for each class. Five different classification models are trained by fine-tuning BERT-based models: four based on gBERT (german BERT) and the remaining one based on a RoBERTa model which has been fine-tuned on Norwegian and Danish economic activity descriptions. The models are then trained in several combinations of keywords, descriptions from the explanatory notes, and the LLM-generated synthetic data. Models which are trained with LLM-generated synthetic data (either in addition to other data or just with synthetic data) outperform the model trained only using keywords, with the model trained just on LLM-generated synthetic descriptions of economic activities being the best one in Top 1 class accuracy and F1-score. In the conclusions, the authors mention that this is a promising approach to automatic codification and note that it makes sense not to just focus on the most suitable class but to determine the top 5 classes. They also remark on the necessity of further testing the models with real-world data, since they only have synthetic data and keywords available for testing.

In [Lim, 2025], a model for the Singapore occupations statistical classification (SSOC) is presented. The model is built with two layers: the first one transforms text to embedding with a BERT-like model (*bge-large-en-v1.5*) and, after that, a “classical” ML classifier produces the final output.

Synthetic data is generated to enhance the SSOC coder performance. The generation procedure involves developing a structured prompt that provides instructions to OpenAI’s *ChatGPT-4o-mini* to create job title and job description pairs based on official definitions from the SSOC 2020 codebook. The prompt specifies instructions to ensure the generated data remains distinct within each detailed SSOC occupation, closely adhering to official definitions, and maintains realistic job descriptions without overstating or understating the scope of responsibilities. For SSOC codes ending with '9' (i.e., occupations that are 'not elsewhere classified' (n.e.c.)), details on the other SSOC codes beginning with the same 4-digits are provided to ensure that the synthetic data for n.e.c. codes does not overlap with other occupations within the same unit group.

The approach produces significant improvements over the basic implementation without generating synthetic data. The addition of these samples reduced the imbalance in the distribution of training data across SSOC classes. Moreover, the results demonstrated that

incorporating synthetic data led to a substantial increase in model performance across all metrics. This highlights the effectiveness of the approach over conventional class-balancing methods, which often involve trade-offs in performance.

Statistics Spain has developed CodIA [Pérez-Bote et al., 2025], a codifier for the Spanish version of the NACE classification, also using the dataset generation-based learning approach. In essence, an LLM is used to generate a synthetic dataset via a sophisticated prompt that utilizes information of the explanatory notes of the classification, and then this synthetic dataset is used, combined with a real-world dataset in which classical data augmentation techniques have been applied, to train a light classification model (fastText). By combining real-world data with the synthetic dataset generated by an LLM, an important improvement in the performance of the fastText model is observed. Moreover, recent experiments using BERT-based models instead of fastText show a promising improvement in performance by using synthetic datasets.

It is also worth noting a finding during the development of CodIA for the NACE classification task: while the best results so far have been obtained by the combination of real and synthetic datasets, the performance of the models trained with only generated data competes and slightly surpasses the performance of models trained with only real data. These results contrast to similar experiments reviewed in the literature: in the vast majority of them, supervised training with real datasets is the best strategy by a large margin. This difference supports the hypothesis that coding for statistical classification is a singular task, where available real data shows specific quality issues and generated samples exhibit outstanding variety and detail due to the availability of titles and explanatory notes.

Finally, it is important to note that the use of synthetic datasets allows us to provide a zero-shot (or few-shot) approach to codifiers for statistical classifications, since it is possible to generate a synthetic dataset of good quality just by using the explanatory notes of the classifications, thus bypassing the need for collecting real-world data and thus greatly reducing the cost of implementing a ML codifier.

Conclusions

In this literature review, we have summarized some of the approaches proposed in the NLP literature for generating synthetic data, especially using LLMs, and devoting special attention to the dataset generation-based learning approach in which an LLM-generated synthetic dataset is used to train a lighter classification model. We have also explored several aspects of synthetic data such as generation workflows and quality evaluation, as well as alternative approaches in the use of LLMs. Finally, we have noted the potential of these techniques in official statistics.

As is apparent from the publication dates of most of the papers, this is a highly active field of research, and there is still lot of exploration to be done in order to achieve optimal results using these techniques. In official statistics these techniques are just starting to be explored and applied, but we believe they have the potential to provide a powerful tool for automatic classification.

6. Bibliography

[Zhou et al., 2024] Zhou, Y., Guo, C., Wang, X., Chang, Y., & Wu, Y. (2024). *A survey on data augmentation in large model era*. arXiv preprint arXiv:2401.15422.

[Bayer et al., 2022] Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). *A survey on data augmentation for text classification*. ACM Computing Surveys, 55(7), 1-39.

[Wang et al., 2024] Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C. T., ... & Zhou, Y. (2024). *A comprehensive survey on data augmentation*. *arXiv preprint arXiv:2405.09591*.

[Ye et al., 2022] Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., & Kong, L. (2022). ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11653–11669). Association for Computational Linguistics.

[Meng et al., 2022] Meng, Y., Huang, J., Zhang, Y., & Han, J. (2022). Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. *arXiv preprint arXiv:2202.04538*.

[Ye et al., 2022] Ye, J., Gao, J., Wu, Z., Feng, J., Yu, T., & Kong, L. (2022). ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3671–3683.

[Gao et al., 2023] Gao, J., Pi, R., Lin, Y., Xu, H., Ye, J., Wu, Z., Zhang, W., Liang, X., Li, Z., & Kong, L. (2023). *Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning*. En *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda.

[Zou et al., 2024] Zou, T., Liu, Y., Li, P., Zhang, J., Liu, J., & Zhang, Y.-Q. (2024). FuseGen: PLM Fusion for Data-generation based Zero-shot Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 2172–2190). Association for Computational Linguistics.

[Long et al., 2024] Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., Wang, H. *On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey* (2024). arXiv preprint arXiv:2406.15126

[Wang et al., 2021] Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M. *Want to Reduce Labeling Cost? GPT-3 Can Help* (2021). Findings of the Association for Computational Linguistics: EMNLP 2021, p. 4195–4205

[Chung et al., 2023] Chung, J., Kamar, E., & Amershi, S. *Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions* (2023). arXiv preprint arXiv:2306.04140.

[Gandhi et al., 2024] Gandhi, S., Gala, R., Viswanathan, V., Wu, T., Neubig, G., *Better Synthetic Data by Retrieving and Transforming Existing Datasets* (2024). arXiv preprint arXiv:2404.14361

[Divekar et al., 2024] Divekar, A., & Durrett, G. *SynthesizRR: Generating Diverse Datasets with Retrieval Augmentation* (2024). arXiv preprint arXiv:2405.10040. Zhou, Y., Guo, C., Wang, X., Chang,

[Li et al., 2023] Li, Z., Zhu, H., Lu, Z., Li, M. *Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations* (2023). Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, p. 10443–10461

[Liu et al., 2024] Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D., & Dai, A. M., *Best Practices and Lessons Learned on Synthetic Data* (2024). arXiv preprint arXiv:2404.07503

[Heß, 2024] Heß, G., *Use of a large language model to derive the economic sector of businesses from unstructured text on economic activities* (2024). Conference on Foundations and Advances of Machine Learning in Official Statistics, Wiesbaden.

[Lim et al., 2025] Lim J., Ng L., Erh R., *Automating Classification with DOS Intelligent Classification Engine (DICE)* (2025). Generative AI and Official Statistics Workshop 2025, Geneva.

[Pérez-Bote et al., 2025] Pérez-Bote, A., Sáez-Calvo, C., Salgado, D., *An ML-based automatic coding machine as an integral element in an overall strategy for NACE implementation* (2025). New Techniques and Technologies for Statistics 2025, Book of Abstracts p. 75–79