Towards Enhancing Deep Active Learning with Weak Supervision and Constrained Clustering

Matthias Aßenmacher[1,2,*], Lukas Rauch[3], Jann Goschenhofer[1,2,4], Andreas Stephan[5,6], Bernd Bischl[1,2,4], Benjamin Roth[5] and Bernhard Sick[3]

[1]Department of Statistics, LMU Munich, Ludwigstr. 33, D-80539 Munich, Germany
[2]Munich Center for Machine Learning (MCML), LMU Munich, Germany
[3]University of Kassel, Wilhelmshöher Allee 73, D-34121 Kassel, Germany
[4]Fraunhofer IIS, Erlangen, Germany
[5]Research Group Data Mining and Machine Learning, University of Vienna, Kolingasse 14-16, A-1090 Vienna
[6]UniVie Doctoral School Computer Science, Vienna, Austria

Abstract

Three fields revolving around the question of how to cope with limited amounts of labeled data are Deep Active Learning (DAL), deep Constrained Clustering (CC), and Weakly Supervised Learning (WSL). DAL tackles the problem by adaptively posing the question of which data samples to annotate next in order to achieve the best incremental learning improvement, although it suffers from several limitations that hinder its deployment in practical settings. We point out how CC algorithms and WSL could be employed to overcome these limitations and increase the practical applicability of DAL research. Specifically, we discuss the opportunities to use the class discovery capabilities of CC and the possibility of further reducing human annotation efforts by utilizing WSL. We argue that the practical applicability of DAL algorithms will benefit from employing CC and WSL methods for the learning and labeling process. We inspect the overlaps between the three research areas and identify relevant and exciting research questions at the intersection of these areas.

Keywords

deep active learning, constrained clustering, weak labels, weak supervision, labels, deep learning

1. Introduction

In the age of large (partially) unlabeled data, we observe a natural rise in research efforts exploring approaches that address situations of missing, low-informative, or incomplete labels; heavily skewed label distributions; or availability of only a few labeled instances. One of the most prominent examples is Deep Active Learning (DAL), which selects unlabeled observations for labeling that are deemed most beneficial for the current model. However, a notable number of benchmark scenarios in this area are partially constructed or artificial and rely on unrealistic assumptions – e.g., perfect knowledge about the true number of classes, or the presence of

IAL@ECML-PKDD'23: 7th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 22nd, 2023, Torino, Italy
*Corresponding author.
matthias@stat.uni-muenchen.de (M. Aßenmacher)
0000-0003-2154-5774 (M. Aßenmacher); 0000-0002-6552-3270 (L. Rauch); 0000-0002-1251-459X (J. Goschenhofer); 0000-0003-0294-9124 (A. Stephan); 0000-0001-6002-6980 (B. Bischl); 0000-0003-0362-0267 (B. Roth); 0000-0001-9467-656X (B. Sick)

omniscient human annotators [1]. This leads to unreliable results, casting doubt on the broader applicability of DAL in practice [2, 3]. Further, incorporating concepts from (a) Constrained Clustering (CC [4, 5]), which uses binary constraints as labels, and (b) Weakly Supervised Learning (WSL), which employs different label-generating mechanisms, may alleviate this limitation and enhance the practical relevance of research in this field. When combining these approaches, we argue that they might complement each other, correct some of the shortcomings of DAL, and mutually extend their capabilities. In the following, we will elaborate on some of the weaknesses of current DAL research, show how CC and WSL potentially alleviate these shortcomings by allowing for dynamic changes of the label set and more efficient learning, and deduce effective research questions. These research questions mainly focus on how pool-based DAL can accommodate a time-varying number of classes via CC and how WSL can support generating the constraints in a meaningful manner.

2. Current Practice of DAL Research

Assuming that a large pool of unlabeled data is available and the goal is to strategically select a subset that is expected to yield the highest performance gains, pool-based DAL can be seen as a scenario of limited labeled data. However, the presumption that the allegedly omniscient oracle is queried for instance-wise labels implies that the number of classes constituting the label set is static and known. We challenge both of these assumptions, since they impose a high barrier to the practical applicability of DAL and severely limit the transferability of research findings into practical settings. As opposed to benchmark data sets with a known, fixed label set and data from the same domain, in real-world settings, it will be much more common to face an unknown and/or dynamic set of classes and samples potentially subject to domain shift [6, 7]. Further critical issues in evaluating DAL include the arbitrary selection of hyperparameters in empirical studies, due to the nature of DAL as a one-time learning problem [8]. Resorting to default hyperparameter values – which are expected to work out-of-the-box without a validation set – limits the comparability across studies as well as the practical implications of these findings [3]. While we do not address this issue, it must be considered an important future challenge.

3. Leveraging Constrained Clustering and Weak Supervision

Instead of using conventionally labeled samples (x_i, y_i) where $y_i \in \mathcal{Y}$, $\mathcal{Y} = \{1, \dots, C\}$, in CC, we consider constraint pairs of the form (x_i, x_j, c_{ij}) where x_i, x_j are two feature vectors and $c_{ij} \in \{0, 1\}$ is the associated binary constraint describing whether the samples are in the same cluster ($c_{ij} = 1$, *Must-Link*) or different clusters ($c_{ij} = 0$, *Cannot-Link*) [9]. Dedicated loss functions allow the training of neural-network-based models that are usable for the inductive prediction of instance-wise class labels [4, 5]. As a rather young research field, characteristics of deep CC are not yet fully explored. For instance, it suffers from constraint imbalancedness in scenarios where an increasing number of underlying distinct clusters (the cardinality $|\mathcal{Y}|$) naturally leads to more Cannot-Link than Must-Link constraints, with a Must-Link proportion of roughly $\frac{1}{|\mathcal{Y}|}$. Further, deep CC shows inferior performance compared to standard Supervised Learning approaches when given a subset of all potential constraint annotations [10, 11] and requires the Hungarian Matching algorithm for model evaluation [12], which is typical for deep clustering. Nevertheless, DAL might benefit from CC's adaptive capabilities by overcoming the assumption of a static label set and simultaneously helping alleviate the problem of imbalancedness in CC. WSL is a promising approach that leverages noisy, incomplete, or imprecise labels based on weak supervision signals, often formalized by so-called labeling functions $\lambda : \mathcal{X} \rightarrow \mathcal{Y} \cup \{\emptyset\}$, where $\emptyset$ means that the labeling function abstains from labeling. Examples include dictionary lookups, heuristics, or pre-trained models [13]. However, there are also several shortcomings and challenges in current WSL research. While WSL allows for cheap and fast label generation, it sometimes struggles to generalize to unseen or out-of-distribution data, since the weak supervision signals do not fully cover the variations present in the data [14]. Similarly to common DAL research, weak labels are typically provided at the instance level. However, in many real-world scenarios, they may be available at different granularities [15]. Finally, while the focus typically lies on classification tasks with a fixed set of predefined classes, this is often unrealistic as the set of classes may evolve over time. While a weak supervision signal typically refers to a single class, it has the potential to be dynamically adjusted over time.

4. Resulting Research Questions

We identify four promising research questions in the intersection between the three research areas to improve the practical applicability of DAL, highlighting commonalities and challenges.

Is it possible to extend DAL beyond known classes via CC? Given the adaptive capabilities of CC models [10], they might be well-suited to extend the practical applicability of DAL research. The authors [10] show how to use CC to detect the true number of classes and to dynamically adapt to a changing number of classes over time. In the world of pool-based DAL, this means that instead of querying the oracle for the label of *one* instance, the desired annotations would be pairwise constraints. This procedure would (a) require less domain knowledge from an annotator's point of view [16] and (b) allow starting the DAL cycle without prior knowledge of the true number of classes, hence resulting in a more realistic setting. However, a problem is the increased computational complexity of the selection: instead of performing the inference step for all n samples, the complexity now scales quadratically with the number of samples. Given that even the computational overhead of performing the inference steps on all n samples sometimes may be prohibitively large, this issue poses a severe threat to feasibility. As a result, dedicated research is needed for active sample selection to address this limitation. Further, existing DAL query strategies would have to be adapted to be capable of executing these pairwise queries (for which they currently do not work out-of-the-box). This can be seen as a chance to develop query strategies that are able to "grasp" the entire space of the data.
In order to answer this research question, it might be beneficial to also closely inspect the methods introduced for the related, but not identical task of (active) domain adaptation.[1] In the realm of DAL for computer vision there has been prior work on (active) domain adaptation [17, 18, 19] while, to the best of our knowledge, no comparable research exists for natural language

[1]Note, that we do not assume a complete shift of the domain, but rather address the scenario of only partial knowledge about, e.g., the true number of classes or the possibility of newly emerging ones for a given domain.

processing. Further work on class distribution mismatch [20] and open-set classification [21] in the computer vision domain could serve as a blueprint to disseminate and adapt these strategies to other domains.

Can we generate constraints in an informed manner? Since WSL methods enable the automatic generation of weak or noisy labels, they represent a promising technique to further reduce human labor in the DAL cycle [22, 23]. While there are established and effective frameworks for WSL on an instance level [13, 24], the generation of weak pairwise constraints is yet to be explored. We argue – especially for a large set of classes – that it might be easier to write appropriate labeling functions λ for generating constraints than for generating instance-wise labels. Very simplistic ideas in the natural language processing domain could be to consider word overlap between documents or cosine similarities of the embeddings, which does not require any domain knowledge (as opposed to writing labeling functions for single instances). This is based on the same intuition, reasoning that it is also an easier task for humans to annotate this way. Yet, such automatic labeling of pairs entails typical WSL problems, such as noisy labels or limited coverage. Therefore, specialized treatment and research are needed. Another approach for, e.g., noisy annotation of texts is the use of the impressive generative capabilities of large language models (e.g. ChatGPT [25]) to generate a suitable second sample x_j, given a pivotal sample x_i and the desired type of constraint c_{ij}. An example of how this approach would look is shown in Figures 1, 2, and 3 in Appendix A. Work in that direction might also be able to solve the problem of increased computational complexity when combining DAL with CC, since only the pivotal sample x_i would have to be selected. Again, sample selection and prompt generation need a dedicated research effort. Moreover, using such a generative approach, it is possible to specifically target classes that are difficult for the model to distinguish.

Does DAL work with mixed supervision? Human-in-the-loop approaches for weak labels beyond instance-wise class labels allow for the combination of labels of different strengths, an area that is referred to as mixed supervision and is connected to multi-task learning [26, 27]. In mixed supervision, annotations for the same data set with different degrees of informativeness are combined for model training, e.g. the mixing of some segmentation masks and some class labels in medical imaging [28]. A first step in that direction would be the design of a DAL query strategy that intelligently queries weak constraint annotations, instance-wise class labels, or both.[2] This could enable a model-driven balancing of the tradeoff between ease of annotation (i.e., quantity, as pairs in some cases[3] might be more effortless to annotate than single instances) and the annotation informativeness (i.e., quality, as instance-wise class labels contain more information than binary constraints). This tradeoff would also allow accounting for the different annotation costs of instance- vs. pairwise labels, where the latter is often assumed to be cheaper.[4]

Can DAL help to break the constraint imbalancedness? A major problem of CC is the increasing imbalancedness of constraint labels with increasing class cardinality: the more classes

[2]A related approach is ALICE [29], with queries aiming at differentiating between classes instead of labeling instances.
[3]This might, e.g., hold for the annotation of pairs short texts or pairs of images for easy object detection tasks.
[4]However, empirically determining the cost ratio of instance- vs. pairwise labels is still an open research question.

exist, the more likely CC is to face a Cannot-Link pair during the data annotation process. In turn, this leads to a notably higher number of Cannot-Link constraints than Must-Link constraints in the training data, consequently hampering the learning process of the model. When enabled to actively select which data samples (or pairs of data samples) to annotate next, this selection process could be steered in a direction where the DAL strategy is used to actively query a higher portion of similar pairs, resulting in a higher likelihood to obtain Must-Link constraints.

5. Conclusion and Outlook

With this contribution, we hope to provide some inspiration for how to potentially combine multiple promising research fields for learning from limited labeled data. When used in conjunction, these methods might be able to alleviate the shortcomings that emerge when each method is used in isolation. This approach may also overcome problems regarding computational overhead as well as limitations caused by insufficient amounts of humanly labeled data sets.

Acknowledgments

We would like to thank the three anonymous referees for their constructive comments and suggestions, which helped improve this paper considerably. This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581.

References

[1] M. Herde, D. Huseljic, B. Sick, A. Calma, A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification, IEEE Access 9 (2021) 166970–166989. URL: https://ieeexplore.ieee.org/abstract/document/9650877. doi:10.1109/ACCESS.2021.3135514.

[2] C. T. Lüth, T. J. Bungert, L. Klein, P. F. Jaeger, Toward Realistic Evaluation of Deep Active Learning Algorithms in Image Classification, CoRR (2023). doi:10.48550/arXiv.2301.10625.

[3] L. Rauch, M. Aßenmacher, D. Huseljic, M. Wirth, B. Bischl, B. Sick, ActiveGLAE: A Benchmark for Deep Active Learning with Transformers, 2023. Accepted at ECML-PKDD 2023.

[4] Y.-C. Hsu, Z. Kira, Neural network-based clustering using pairwise constraints, arXiv preprint arXiv:1511.06321 (2015).

[5] Y. Hsu, Z. Lv, J. Schlosser, P. Odom, Z. Kira, Multi-class classification without multi-class labels, 2019.

[6] K. Cao, M. Brbic, J. Leskovec, Open-world semi-supervised learning, 2021. URL: https://openreview.net/forum?id=6VhmvP7XZue.

[7] M. N. Rizve, N. Kardan, S. Khan, F. Shahbaz Khan, M. Shah, Openldn: Learning to discover novel classes for open-world semi-supervised learning, in: Computer Vision–ECCV 2022:

17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI, Springer, 2022, pp. 382–401.

[8] D. Kottke, A. Calma, D. Huseljic, G. Krempl, B. Sick, et al., Challenges of reliable, realistic and comparable active learning evaluation, in: Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning, 2017, pp. 2–14.

[9] H. Zhang, T. Zhan, S. Basu, I. Davidson, A framework for deep constrained clustering, Data Mining and Knowledge Discovery 35 (2021) 593–620.

[10] J. Goschenhofer, P. Ragupathy, C. Heumann, B. Bischl, M. Aßenmacher, CC-top: Constrained clustering for dynamic topic discovery, in: Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 26–34. URL: https://aclanthology.org/2022.evonlp-1.5.

[11] J. Goschenhofer, B. Bischl, Z. Kira, Constraintmatch for semi-constrained clustering, International Joint Conference on Neural Networks (IJCNN) (2023).

[12] H. W. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly 2 (1955) 83–97.

[13] A. J. Ratner, S. H. Bach, H. R. Ehrenberg, C. Ré, Snorkel: Fast training set generation for information extraction, in: Proceedings of the 2017 ACM international conference on management of data, 2017, pp. 1683–1686.

[14] J. Zhang, C.-Y. Hsieh, Y. Yu, C. Zhang, A. Ratner, A survey on programmatic weak supervision, 2022. arXiv:2202.05433.

[15] C. Shin, W. Li, H. Vishwakarma, N. C. Roberts, F. Sala, Universalizing weak supervision, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=YpPiNigTzMT.

[16] Y. Fu, B. Li, X. Zhu, C. Zhang, Active learning without knowing individual instance labels: A pairwise label homogeneity query approach, IEEE Transactions on Knowledge and Data Engineering 26 (2014) 808–822. doi:10.1109/TKDE.2013.165.

[17] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, G. Wang, Active learning for domain adaptation: An energy-based approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 8708–8716.

[18] V. Prabhu, A. Chandrasekaran, K. Saenko, J. Hoffman, Active domain adaptation via clustering uncertainty-weighted embeddings, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8505–8514.

[19] B. Fu, Z. Cao, J. Wang, M. Long, Transferable query selection for active domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7272–7281.

[20] P. Du, S. Zhao, H. Chen, S. Chai, H. Chen, C. Li, Contrastive coding for active learning under class distribution mismatch, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8927–8936.

[21] K.-P. Ning, X. Zhao, Y. Li, S.-J. Huang, Active learning for open-set annotation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 41–49.

[22] L. Rauch, D. Huseljic, B. Sick, Enhancing active learning with weak supervision and transfer learning by leveraging information and knowledge sources, Proceedings of the

6th Workshop on Interactive Adaptive Learning (2022).
[23] B. Boecking, W. Neiswanger, E. Xing, A. Dubrawski, Interactive weak supervision: Learning useful heuristics for data labeling, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=IDFQI9OY6K.
[24] A. Sedova, A. Stephan, M. Speranskaya, B. Roth, Knodle: Modular weakly supervised learning with PyTorch, in: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), Association for Computational Linguistics, Online, 2021, pp. 100–111. URL: https://aclanthology.org/2021.repl4nlp-1.12. doi:10.18653/v1/2021.repl4nlp-1.12.
[25] OpenAI, Chatgpt: Optimizing language models for dialogue, 2022. URL: https://openai.com/blog/chatgpt/, accessed: 2023-01-10.
[26] M. Crawshaw, Multi-task learning with deep neural networks: A survey, arXiv preprint arXiv:2009.09796 (2020). URL: https://arxiv.org/abs/2009.09796.
[27] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, E. Fetaya, Multi-task learning as a bargaining game, International Conference on Machine Learning (2022). URL: https://arxiv.org/pdf/2202.01017v2.pdf.
[28] B. Liu, C. Desrosiers, I. B. Ayed, J. Dolz, Segmentation with mixed supervision: Confidence maximization helps knowledge distillation, Medical Image Analysis 83 (2023) 102670.
[29] W. Liang, J. Zou, Z. Yu, ALICE: Active learning with contrastive natural language explanations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4380–4391. URL: https://aclanthology.org/2020.emnlp-main.355. doi:10.18653/v1/2020.emnlp-main.355.
[30] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in neural information processing systems 28 (2015).

A. Generating constraints with ChatGPT

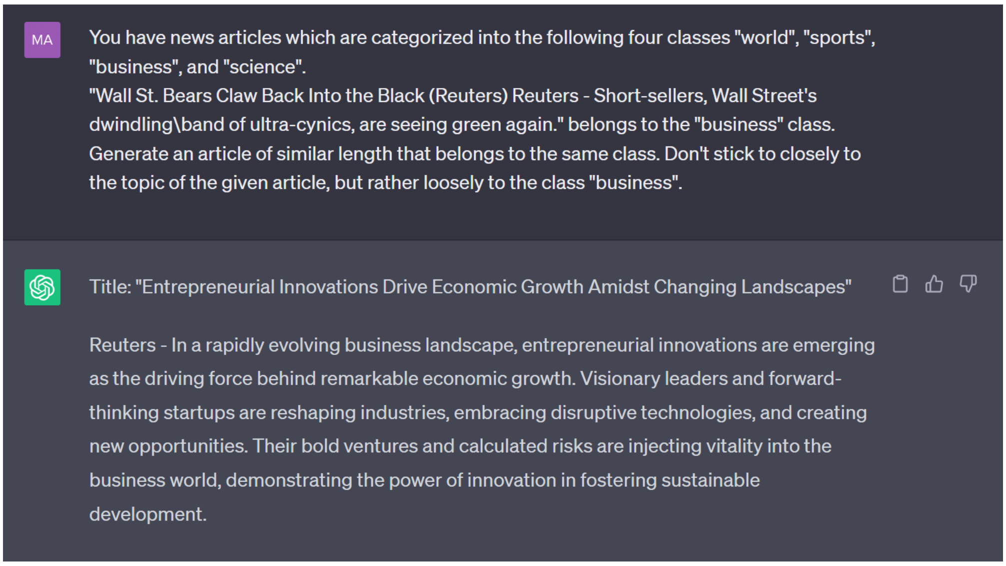

Figure 1: An example (based on the AG News data set [30]) for generating the second sample for a constraint pair (Must-Link) using a large langauge model.

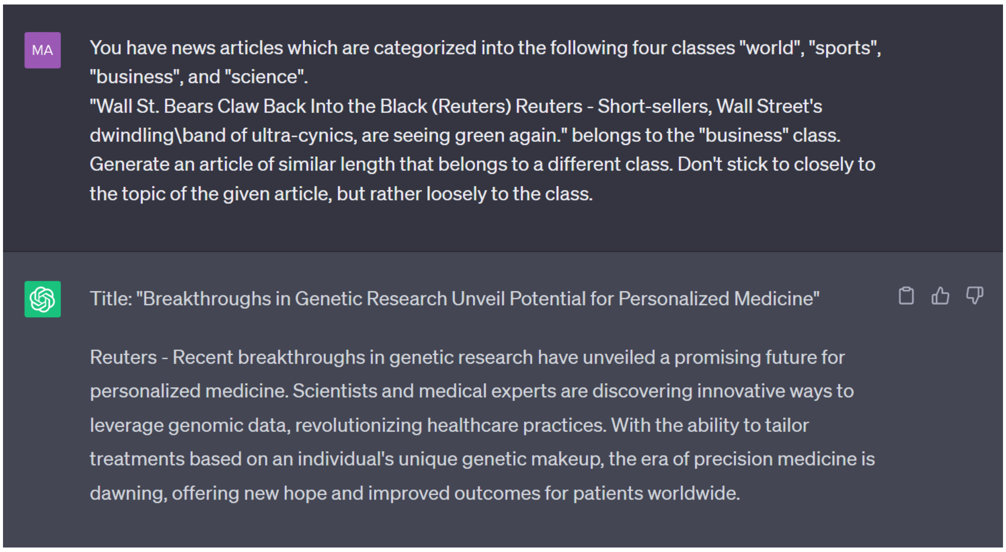

Figure 2: An example (based on the AG News data set [30]) for generating the second sample for a constraint pair (Cannot-Link) using a large language model.

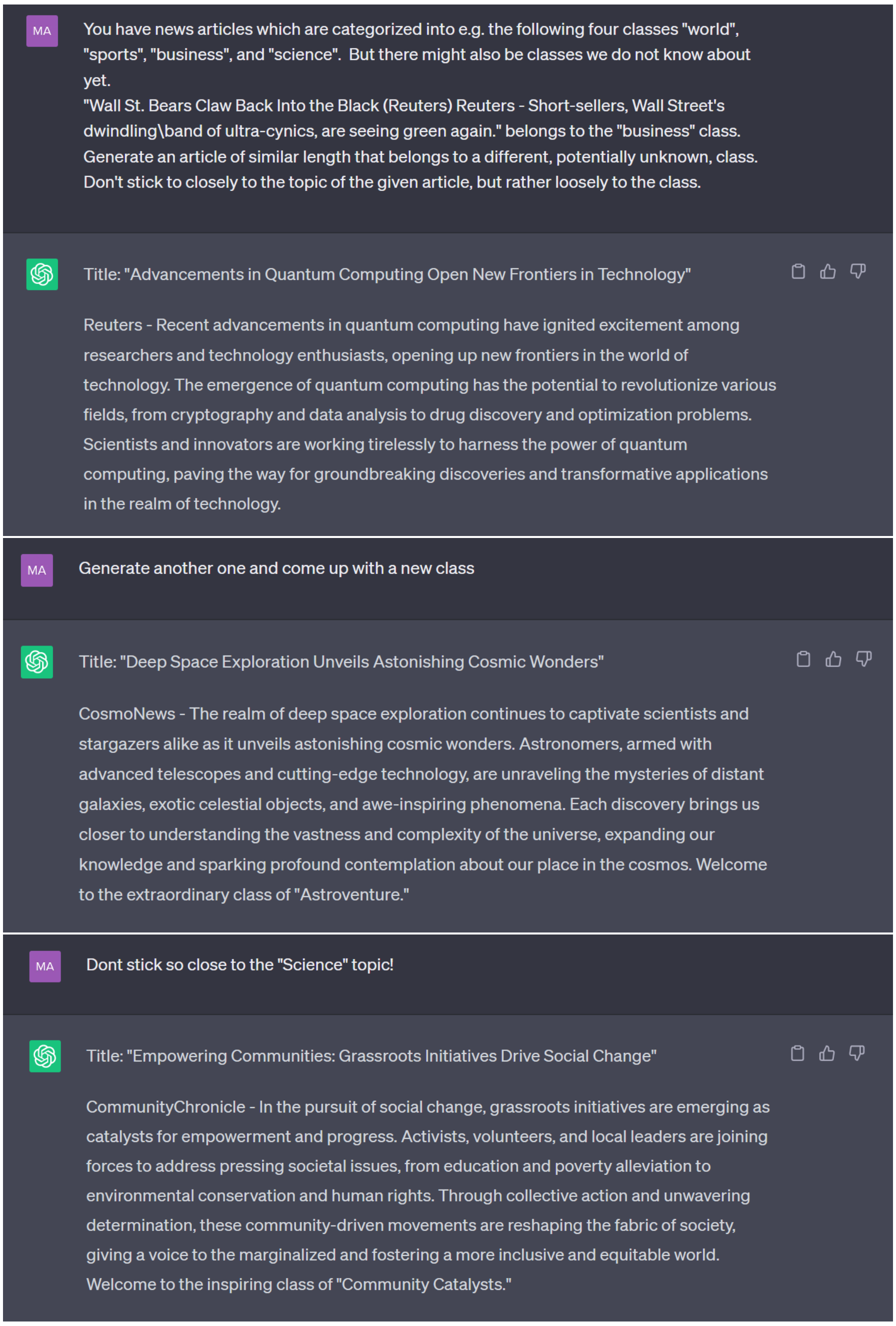

Figure 3: Generating samples belonging to an unseen class proves to be more challenging, but still seems possible. Please note, that these experiments were not performed in a scientifically rigorous manner, but rather just some first tentative experiments via trial and error.