

DSBDA Template: The Name of the Title is Hope

<https://tinyurl.com/dsbda-template>

Ansgar Scherp
ansgar.scherp@uni-ulm.de
Ulm University
Ulm, Germany

Ben Trovato*
G.K.M. Tobin*
trovato@corporation.com
webmaster@marysville-ohio.com
Institute for Clarity in
Documentation
Dublin, Ohio, USA

Lars Thørvæld
The Thørvæld Group
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
San Antonio, Texas, USA
cpalmer@prl.com

John Smith
The Thørvæld Group
Hekla, Iceland
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
New York, USA
jpkumquat@consortium.net

ABSTRACT

This template is for papers, research-based group work reports, seminar works etc. It is based on a common ACM style, which is both popular in the computer science research community as well as well maintained.

Note on the use of ChatGPT: We are following the procedure of the International Conference on Machine Learning (ICML), which states: “The Large Language Model (LLM) policy for ICML 2023 prohibits text produced entirely by LLMs (i.e., “generated”). This does not prohibit authors from using LLMs for editing or polishing author-written text.”. Source: <https://icml.cc/Conferences/2023/llm-policy>.

For comments and feature requests, please email Ansgar at ansgar.scherp@uni-ulm.de.

Submission: *We pledge to make the source code and additional resources publicly available upon acceptance of the paper. An (anonymous) preview for the reviewers can be found at: <http://anonymo.us/me>.*

Submission (if already available on arXiv): *An earlier version of this paper has been published on arXiv (add cite). We release the source code upon acceptance of the paper.*

Final: *The source code and additional resources are available at: <http://anonymo.us/me>*

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems; Redundancy; Robotics**; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

1 INTRODUCTION

For the author information. Create an ORCID and add it to your record, see example of first author. You can obtain an ORCID here: <https://orcid.org/>

Note: This template is based on the official “Association for Computing Machinery (ACM) - SIG Proceedings Template” provided on Overleaf. A documentation is provided in this project. The template is taken from Overleaf: <https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sig-proceedings-template/bmvfhcdnxfty>

The official URL to this Overleaf template is: <https://www.overleaf.com/latex/templates/dsbda-templateforpaper-annotated/svwvwwqxfxtp> You may also use the view link (ready only): <https://www.overleaf.com/read/mpmsdhfcwdfk>.

If you look for a template for presentations/slides, Fabian Singhofer is so kind to share his for DSBDA: <https://www.overleaf.com/read/qxrdtnzrprwc>

Links are “read”-links, so one can copy it into a new project. By default, the language is set to American English.

The concept of the teaching programme is also documented and available here: <https://github.com/data-science-and-big-data-analytics/teaching-examples/blob/main/Scherp-TdL21-vortrag.pdf>

Note that there are also new writing tools that support academic writing. For example, Grammarly: <https://www.grammarly.com/blog/academic-writing/>

Note: Yellow boxes provide background information, additional notes, recommendations, etc. and can later be removed.

Apply Jennifer Widom structure, which is encoded here in the yellow boxes.

What is the motivation?

Motivate your work.

What is the problem?

Describe in precise terms what the problem is that you address. This definition of the problem is used/referred to throughout the paper.

Why is it a problem?

Describe the relevancy of the problem.

Why is it not yet solved?

For the abstract, please follow the Jennifer Widom structure.

*Both authors contributed equally to this research.

Describe why are existing solutions insufficient.

What is our solution approach?

Describe the method/algorithm that you propose to solve the problem.

What are the results?

Describe key results from your experiments. Mention datasets, measures, and observations. Reflect on the key insights by a brief discussion. Make the reader interested in your paper.

What is your contribution?

Below, we summarize our contributions.

Point of discussion: your contributions list.

- Provide a bullet-itemized list of research questions that you address.
- Later, each research question will then be turned into a contribution, i. e., a brief answer to the question is given.

The remainder of the paper is organized as follows. Below, we summarize the related works. Section 3 provides a problem statement and introduces our models/methods. The experimental apparatus is described in Section 4. An overview of the achieved results is reported in Section 5. Section 6 discusses the results, before we conclude.

Instructions

Follow the structure suggested above. Write explicit paragraphs for each of the questions. Furthermore, make sure that the introduction picks up every statement made by the abstract. The goal of the introduction is to extend the gist provided by the abstract by giving more detail, more context, explanations, and, very important, citations to definitions, related work, and methods.

2 RELATED WORK

When reading the related work, we aim to understand the method(s), datasets used, results of the experiments, and what the results mean, i. e., how the authors argue about the results in the discussion.

Instructions

To check the trustworthiness of results, we perform always some checks (derived from [1]). Papers, where one has to tick one of the items below, do not allow for a fair comparison with the state of the art. Reasons include that they

- used different or non-standard benchmark datasets,
- modified the datasets to use a different number of classes (i. e., reducing the number of classes in the preprocessing),
- modified the datasets to use additional information (e. g., additional header metadata in the 20ng text dataset),
- employed different train-test splits (e. g., use more training samples than others),
- used a different, smaller number of training examples (e. g., run their methods only on 5% of the training data while using a benchmark dataset),
- not report the train-test splits (and thus the training data used remains unclear),
- do not report hyperparameter values (particularly the learning rate),
- do not report an average over multiple runs of the experiments together with the standard deviation (Avg. and SD will allow to assess the influence of random factors like the initialization of model weights),
- have not optimized or do not use optimal hyperparameter values (e. g., the learning rate strongly influences the results as demonstrated at the examples of BERT and RoBERTa by Galke et al. [1]),
- do unusual preprocessing on the datasets (e. g., apply preprocessing for models that do not require it like BERT, drop samples in a multi-labeling task that have 1 label and thus modify the datasets, etc.), or
- are unclear about the measure(s) used (e. g., while writing “we use the F-score” most likely means the (harmonic) F1-score, it still does not detail if micro-averaging, macro-averaging, or samples-averaging F1 is reported).

The rationales for not using benchmark datasets or employing other train-test splits are not always clear. Also, the papers often do not properly report hyperparameter values or miss reporting any other of the items above.

As a general rule when reading related work

Be suspicious and ask yourself: “Can I trust their results?” Keep in mind: A primary objective of the paper is to put their method in a good light.

And an important lesson when searching for literature.

Lesson learned (once) again!

If you search for literature and do not find anything. Likely you just did not search for the right keywords. For example, if you search for research on "(source) code segmentation", you will be disappointed (or happy) not to find any. But do not be a fool. There is work, it is "text segmentation" a classical area in natural language processing. You just have to think about source code being an (artificial) language that any modern tool will process in the same way as a natural language. A good hint is also if the task is visible in the community. For text segmentation there exists its own category on Papers with Code, see <https://paperswithcode.com/task/text-segmentation>.

Writing hint: Use [?] or ?].
But always put a tilde (~) before the \cite.

2.1 Area 1

2.2 Area 2

2.3 Area ...

2.4 Summary/Reflection

What do we learn from the literature concerning your work? Where are their strengths, and where are their weaknesses? What is different in the related work compared to the proposed approach?

3 METHODS [OR MODELS]

Methods : Which methods do apply?

3.1 [Problem Statement/Problem Formalization]

(if not done as part of the introduction)

3.2 Assumptions

- What are the assumptions that you make?

Note: make sure there is an explicit section or subsection called "Assumptions" in your paper.

3.3 Methods for Aspect 1

Point of Discussion: Provide a bullet-itemized list of the aspects that are considered by your research. For each aspect, provide a description of the methods/models used and proposed (own methods). Make sure it is consistent with the research questions/contributions describe in the introduction.

Example: Aspects are: a) clustering algorithms, b) embedding methods, c) similarity measures. Instances for a) are DBCAN, k-means, etc., b) TF-IDF, BERT, etc., c) cosine similarity.

- Method 1
- Method 2
- ...

3.4 Methods for Aspect 2

3.5 Methods for Aspect 3

3.6 Summary

4 EXPERIMENTAL APPARATUS

Follow the description of the experimental apparatus given the structure below.

Make sure to cover the questions provided in the EMNLP checklist, see Appendix D.

4.1 Datasets

Datasets: Which datasets do you use? Provide descriptive statistics, usually in tabular form.

Point of Discussion: Make sure that your datasets fit to the problem and research questions, respectively. Make sure that the datasets are available. Available means that you have a) the license obtained (if needed) and b) the datasets are actually on your disk (copied).

4.2 Preprocessing OR Pre-processing

4.3 Procedure

Point of Discussion: Describe which methods you use along the aspects defined in your research, on which datasets they are applied, etc. Make sure it reflect fully the experiments that you want to carry out according to your own plan defined in the research questions.

Procedure : How do you run your experiments?

Note: Preprocessing can also be part of procedure.

4.4 Hyperparameter Optimization

Note: If space is limited, this can be moved to supplementary materials

Point of Discussion: What are the (critical) hyperparameters that you need to consider (beyond the learning rate)? How do you plan to optimize the hyperparameters with respect to the models and datasets? What is the hyperparameter search space?

4.5 Measures OR Metrics

Measure: How do you measure the results?

Point of Discussion: Regarding the measurements and what to measure, i. e., to which level of detail, please carefully read: John Ousterhout's article on "Always Measure One Level Deeper" [3].

5 RESULTS

- Report your results in tabular or otherwise structured form.
- Limit to objective results, no interpretation of results

5.1 RQ1 Results

5.2 RQ2 Results

5.3 ... Results

6 DISCUSSION

- Now interpret and reflect on your results.

6.1 Key Scientific Insights [Gained from the Results]

- What is the key takeaway? Reflect on the results (what have we learned from them)?
 - What are the key results of your research?
 - What interesting insights could you obtain?
 - Break down by research question.

6.2 Threat to Validity

- Why may your results be biased/not trustworthy? And why in fact are they trustworthy! How reliable are your analyses? Meaning, critically reflect on whether there may be errors / biases in your analyses. So: What (possible) threats exist that could have made the results unreliable, AND why are these not threats?
 - Trick is to write down potential threats and explain why they don't hold true here!
 - How reliable are your analyses? Meaning, critically reflect on whether there may be errors / biases in your analyses.

6.3 Generalization

- Will the results be transferable/generalize to other datasets, tasks, models, etc?
 - Can one transfer the insights/results to other datasets? ... other scenarios? ... other algorithms? Why can we assume that the results generalize? Why?

6.4 Future Work and Impact

What is future work?

What is the general impact of your work? – pick up arguments from introduction etc.

[- But also: What is the practical impact.]

7 CONCLUSION

Summarize the key results in an interesting and new way. For example by expanding it to a general broader scope of science, economics, impact to life, etc. :-)

Provide a brief outlook to future work! (If not described in Section 6.4)

LIMITATIONS

- Reflect on the limitations of your work, so what conclusion cannot or should not be derived from the work.

See also EMNLP's **Mandatory Discussion of Limitations**.

We believe that it is also important to discuss the limitations of your work, in addition to its strengths. EMNLP 2023 requires all papers to have a clear discussion of limitations, in a dedicated section titled "Limitations". This section will appear at the end of the paper, after the discussion/conclusions section and before the references, and will not count towards the page limit. Papers without a limitation section will be automatically rejected without review.

[...]

While we are open to different types of limitations, just mentioning that a set of results have been shown for English only probably does not reflect what we expect. Mentioning that the method

works mostly for languages with limited morphology, like English, is a much better alternative. In addition, limitations such as low scalability to long text, the requirement of large GPU resources, or other things that inspire crucial further investigation are welcome.

https://2023.emnlp.org/calls/main_conference_papers/#mandatory-discussion-of-limitations

AUTHOR STATEMENT

Author statement based on CRediT (Contributor Roles Taxonomy), see: <https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement>

ACKNOWLEDGMENTS

This template is co-funded under the "2LIKE - Artificial Intelligence for Individualised Learning Path and Processes" (16DHBKI001) project by the German Federal Ministry of Education and Research (BMBF) and the Ministry of Science, Research and the Arts Baden-Württemberg within the funding line Artificial Intelligence in Higher Education.

IF YOU USE THE bwHPC CLUSTER, YOU CAN ADD:
The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

The presented research is the result of a Master module "Project Data Science" taught at the University of Ulm in SEMESTER+YEAR. The last author is supervisor of the student group.¹

The presented research is the result of a Master module "Project Data Science" taught at the University of Ulm in 2022. The last author is supervisor of the student group.

REFERENCES

- [1] Lukas Galke, Andor Diera, Bao Xin Lin, Bhakti Khera, Tim Meuser, Tushar Singhal, and Ansgar Scherp. 2023. Are We Really Making Much Progress in Text Classification? A Comparative Review. *CoRR* abs/2204.03954 (2023). <https://doi.org/10.48550/ARXIV.2204.03954> arXiv:2204.03954
- [2] Gregor Große-Bölting, Chifumi Nishioka, and Ansgar Scherp. 2015. A Comparison of Different Strategies for Automated Semantic Document Annotation. In *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015, Palisades, NY, USA, October 7-10, 2015*, Ken Barker and José Manuel Gómez-Pérez (Eds.). ACM, 8:1–8:8. <https://doi.org/10.1145/2815833.2815838>
- [3] John K. Ousterhout. 2018. Always measure one level deeper. *Commun. ACM* 61, 7 (2018), 74–83. <https://doi.org/10.1145/3213770>

A SUPPLEMENTARY MATERIALS

Note: Backward references to main part of the paper is ok. But do not directly refer to figures or tables from body to here.

A.1 Extended Related Work

A.2 Extended Results

A.3 Hyperparameter Optimization

A.4 Detailed Discussions

A.5 ...

B USEFUL RESEARCH RESOURCES

B.1 Very Interesting Paper

Language Model "Alternatives"

¹Author is contributing Conceptualization, Writing - Review & Editing, and Supervision. Statement is based on the Contributor Roles Taxonomy, see: <http://credit.niso.org/>

- S4: Efficiently Modeling Long Sequences with Structured State Spaces, <https://arxiv.org/abs/2111.00396>
- xLSTM: Extended Long Short-Term Memory, <https://arxiv.org/abs/2405.04517>
- KAN: Kolmogorov-Arnold Networks, <https://arxiv.org/abs/2404.19756>
- gMLP: Pay Attention to MLPs, <https://arxiv.org/abs/2105.08050>
- Pretraining Without Attention, <https://arxiv.org/abs/2212.10544>

Graphs

- Mirage: Model-Agnostic Graph Distillation for Graph Classification, <https://arxiv.org/abs/2310.09486>

Language

- Do Llamas Work in English? On the Latent Language of Multilingual Transformers, <https://arxiv.org/abs/2402.10588>
- Evolutionary Optimization of Model Merging Recipes, <https://arxiv.org/abs/2403.13187>
TL;DR: Language model merging on the level of layers.
- When LLMs are Unfit Use FastFit: Fast and Effective Text Classification with Many Classes, <https://arxiv.org/abs/2404.12365v1>
- Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs, <https://aclanthology.org/2024.eacl-long.5/>

B.2 Interesting Paper

Language

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, <https://arxiv.org/abs/2403.09611>
TL;DR: Show among others that adding automatically generated image captions improves the image classification task.
- Better & Faster Large Language Models via Multi-token Prediction, <https://arxiv.org/abs/2404.19737>
TL;DR: Learn to predict n tokens in causal language modeling rather than a single token is better.

Graphs

- Graph Language Models, <https://arxiv.org/abs/2401.07105>
TL;DR: Transfer of T5's model weights and modification of the attention mechanism to support graphs.

B.3 Surveys

B.3.1 Machine Learning. Attention

* A General Survey on Attention Mechanisms in Deep Learning, <https://arxiv.org/abs/2203.14263>

Continual Learning

* A Wholistic View of Continual Learning with Deep Neural Networks: Forgotten Lessons and the Bridge to Active and Open World Learning, <https://arxiv.org/abs/2009.01797>

* A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning, <https://www.sciencedirect.com/science/article/pii/S089360823000142X>

Distillation

* Knowledge Distillation: A Survey, <https://arxiv.org/abs/2006.05525>

B.3.2 Graphs. * Deep Learning on Graphs: A Survey <https://arxiv.org/abs/1811.04202>

* Comprehensible Artificial Intelligence on Knowledge Graphs: A survey, <https://www.sciencedirect.com/science/article/pii/S1570826821000355>

* Continual Graph Learning: A Survey, <https://arxiv.org/abs/2301.12230>

* A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions, <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00876-4>

* A survey of graph neural networks in various learning paradigms: methods, applications, and challenges, <https://link.springer.com/article/10.1007/978-3-031-1022-10321-2>

* A Comprehensive Survey on Automatic Knowledge Graph Construction, <https://dl.acm.org/doi/10.1145/3618295>

* A Comprehensive Survey on Deep Graph Representation Learning Methods, <https://dl.acm.org/doi/pdf/10.1613/jair.1.14768>

* Graph Self-Supervised Learning: A Survey, <https://arxiv.org/abs/2103.00111>

* A Survey on Temporal Knowledge Graph Completion: Taxonomy, Progress, and Prospects, <https://arxiv.org/abs/2308.02457>

* Position Paper: Challenges and Opportunities in Topological Deep Learning <https://arxiv.org/abs/2402.08871>

* Uncertainty in Graph Neural Networks: A Survey, <https://arxiv.org/abs/2403.02200>

Summarization with GNNs. * A Survey on Graph Condensation, <https://arxiv.org/abs/2402.02000>

* A Comprehensive Survey on Graph Summarization with Graph Neural Networks, <https://arxiv.org/abs/2302.06114>

* A Survey on Extractive Knowledge Graph Summarization: Applications, Approaches, Evaluation, and Future Directions, <https://arxiv.org/abs/2402.12001>

* A Comprehensive Survey on Graph Reduction: Sparsification, Coarsening, and Condensation, <https://arxiv.org/abs/2402.03358>

* A Survey on Structure-Preserving Graph Transformers, <https://arxiv.org/abs/2402.03358>
Structural Summarization

* Structural Summarization of Semantic Graphs Using Quotients, <https://doi.org/10.4230/TGDK.1.1.12>

Link Prediction

* Beyond Transduction: A Survey on Inductive, Few Shot, and Zero Shot Link Prediction in Knowledge Graphs, <https://arxiv.org/abs/2312.04997>

* A Survey on Graph Classification and Link Prediction based on GNN, <https://arxiv.org/abs/2307.00865>

Recommender

* A survey of graph neural network based recommendation in social networks, <https://www.sciencedirect.com/science/article/abs/pii/S0925231321000142X>

* Graph Neural Networks in Recommender Systems: A Survey, <https://arxiv.org/abs/2011.02260>

* A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions, <https://dl.acm.org/doi/full/10.1145/3592222>

Distillation

* Graph-based Knowledge Distillation: A survey and experimental evaluation, <https://arxiv.org/abs/2302.14643>

* Knowledge Distillation on Graphs: A Survey, <https://arxiv.org/abs/2302.00219>

Continual Learning

* Continual Learning on Graphs: Challenges, Solutions, and Opportunities, <https://arxiv.org/abs/2402.11565>

* Graph Learning under Distribution Shifts: A Comprehensive Survey on Domain Adaptation, Out-of-distribution, and Continual Learning, <https://arxiv.org/abs/2402.16374>

B.3.3 Natural Language Processing. * On the Opportunities and Risks of Foundation Models, <https://arxiv.org/abs/2108.07258> [argue among others that due to the huge resources required by language models, research on them is pushed into the hands of a few global industrial players only]

* A Comprehensive Overview of Large Language Models, <https://arxiv.org/abs/2307.06435>

- * Large Language Models: A Survey, <https://arxiv.org/abs/2402.06196>
- * The Life Cycle of Knowledge in Big Language Models: A Survey, <https://arxiv.org/abs/2303.07616>
- * A Survey of Large Language Models, <https://arxiv.org/abs/2303.18223>
- * A Survey of Knowledge Enhanced Pre-Trained Language Models, <https://ieeexplore.ieee.org/document/10234662>
- * A Survey on Deep Semi-Supervised Learning, <https://ieeexplore.ieee.org/document/9941371>
- Efficient LLM
- * Large Language Models: A Survey, <https://arxiv.org/abs/2312.03863> and <https://github.com/AIoT-MLSys-Lab/Efficient-LLMs-Survey>
- Continual
- * Continual Learning for Large Language Models: A Survey, <https://arxiv.org/abs/2402.01364>
- Classification
- * A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?, <https://ieeexplore.ieee.org/document/10380590>
- * Graph Neural Networks for Text Classification: A Survey, <https://arxiv.org/abs/2304.11534>
- * Recent Advances in Hierarchical Multi-label Text Classification: A Survey, <https://arxiv.org/abs/2307.16265>
- * Text classification using embeddings: a survey, <https://link.springer.com/article/10.1007/s10115-023-01856-z>
- * Deep learning, graph-based text representation and classification: a survey, perspectives and challenges, <https://link.springer.com/article/10.1007/s10462-022-10265-7>
- * A Survey of Cross-Lingual Text Classification and Its Applications on Fake News Detection, <https://www.worldscientific.com/doi/10.1142/S2811032323500030>
- Augmentation
- * A Survey on Data Augmentation for Text Classification, <https://dl.acm.org/doi/10.1145/3544558>
- KG Editing with LLM
- * Knowledge Editing for Large Language Models: A Survey, <https://arxiv.org/abs/2310.16218>
- Augmentation
- * Augmented Language Models: a Survey, <https://arxiv.org/abs/2302.07842>
- * Retrieval-Augmented Generation for Large Language Models: A Survey, <https://arxiv.org/abs/2312.10997>
- Interpretability
- * Post-hoc Interpretability for Neural NLP: A Survey, <https://arxiv.org/abs/2108.04840>
- Specific Domains
- * Artificial Intelligence for Literature Reviews: Opportunities and Challenges, <https://arxiv.org/abs/2402.08565>
- Instance Selection
- * A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification, <https://dl.acm.org/doi/10.1145/3582000>
- Question Answering
- * Deep learning-based question answering: a survey, <https://link.springer.com/article/10.1007/s10115-022-01783-5>
- * Modern Question Answering Datasets and Benchmarks: A Survey, <https://arxiv.org/abs/2206.15030>
- * Pretrained Transformers for Text Ranking: BERT and Beyond, <https://arxiv.org/abs/2010.06467>
- * Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges, <https://dl.acm.org/doi/10.1145/3648471>
- * Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting, <https://arxiv.org/abs/2306.17563>
- B.4.1 Time Series Analysis. * Foundation Models for Time Series Analysis: A Tutorial and Survey, <https://arxiv.org/abs/2403.14735v1>

B.5 Books

- B.5.1 *Machine Learning*. Probabilistic Machine Learning: An Introduction <https://probml.github.io/pml-book/book1.html>
- Probabilistic Machine Learning: Advanced Topics <https://probml.github.io/pml-book/book2.html>
- Lifelong Machine Learning <https://www.cs.uic.edu/liub/lifelong-machine-learning.html>
- Lifelong Learning: <https://www.cs.uic.edu/liub/lifelong-machine-learning-draft.pdf>
- The Modern Mathematics of Deep Learning <https://arxiv.org/abs/2105.04026>
- Probabilistic ML intro, <https://probml.github.io/pml-book/book1.html>, Probabilistic ML Adv, <https://probml.github.io/pml-book/book2.html>, Interpretable Machine Learning
- A Guide for Making Black Box Models Explainable <https://christophm.github.io/interpretable-ml-book/>
- B.5.2 *Graphs*. Deep Learning on Graphs, https://yaoma24.github.io/dlg_book/
- Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges <https://arxiv.org/abs/2104.13478>
- GeomDL (also videos etc): <https://geometricdeeplearning.com/>, Graph Representation Learning, https://www.cs.mcgill.ca/~wlh/grl_book/

B.6 Natural Language Processing

- * Eisenstein - Natural Language Processing, <https://cseweb.ucsd.edu/~nnakashole/nov18.pdf> (also: <https://www.amazon.de/Jacob-Eisenstein/dp/0262042843/>)
- * Formal Aspects of Language Modeling, 2023, <https://arxiv.org/abs/2311.04329>
- * Formal Aspects of Language Modeling, 2024, <https://drive.google.com/file/d/1PtxuMe6JZyBXBuuGkgDnnD3JRJE15j/view>
- Training, Fine Tuning, Inference and Applications of Language Models, <https://drive.google.com/file/d/1PtxuMe6JZyBXBuuGkgDnnD3JRJE15j/view>
- Modules
- * Advanced Formal Language Theory, Spring 2023, <https://rycolab.io/classes/aflls23/>
- * Large Language Models, Spring 2023, <https://rycolab.io/classes/llm-s23/>
- Geometric and Computational Geometry <https://www.cs.cmu.edu/afs/cs/afs14/Handouts/BKOS.pdf>

C DATA SCIENCE AND BIG DATA ANALYTICS (DSBDA) GROUP

C.1 Data Science Readings

We are running a reading club on Data Science on Wednesdays.

How it works: Idea of the reading club is to have a joined chat about recent research papers. Particular focus is text analytics and graph analytics, and general recent methods in deep learning.

Procedure is usually as follows:

- Someone proposes a paper/topic, which is well before the meeting disseminated.

- So everyone has time to read the paper and is actually also expected to have read the paper (otherwise discussions are not so much fun!)
- During the meeting, the proposer briefly summarizes the paper, including key strengths and weaknesses.
- Followed by a round-robin quick feedback from everyone.
- Discussion goes into the details ... :-)

How to subscribe: Interested? Go here to subscribe: <https://imap.uni-ulm.de/lists/subscribe/data-science-readings>

This is a mailing list on which you receive current information: <mailto:data-science-readings@lists.uni-ulm.de>

C.2 Lectures, Seminars, Project Groups, and Theses

Lectures: We offer a couple of different lectures for both BSc and MSc students. These are available for self-enrolment with all materials available for download. Please contact us to get information which lectures will be offered the next terms.

- “Graph Analytics and Deep Learning”, Self-enrolment for slides (winter 2022/23): <https://moodle.uni-ulm.de/course/view.php?id=36399>
- “Text Analytics and Deep Learning”, Self-enrolment for slides (winter 2021/22): <https://moodle.uni-ulm.de/course/view.php?id=26119>
- “Web Information Retrieval (and Deep Learning)”, Self-enrolment for slides (summer 2021): <https://moodle.uni-ulm.de/course/view.php?id=22260>
- “Advanced Methods in) Data Mining and Machine Learning”, Self-enrolment for slides (winter 2020/21): <https://moodle.uni-ulm.de/course/view.php?id=16999>
There are also slides for the full 4 SWS module (same moodle course): <https://moodle.uni-ulm.de/mod/folder/view.php?id=254324>

My concept for research-based teaching: https://www.uni-ulm.de/fileadmin/website_uni_ulm/zle/Tag_der_Lehre/downloads/Scherp-TdL21-vortrag.pdf

Seminar and Projects: We also regularly offer seminars on data science (BSc/MSc), as well as the module “Project Data Science”. For projects, please contact us.

Theses: If you are interested in a BSc or MSc thesis, please contact us. We have compiled a couple of topics here: https://docs.google.com/presentation/d/1k1aEZYX_UM8rWlojgGTV11O85Lu104e2Killed/CBDg-k-9A

C.3 Examples of Student Submissions

This folder contains examples of submissions from the last years (in PDF).

<https://github.com/data-science-and-big-data-analytics/teaching-examples>

Please refer to the corresponding sub-folders for an example relevant to a practical group project submitted in the context of a lecture, MSc project, seminar (written for MSc but also suitable for BSc), and MSc thesis.

C.4 Examples of Data Science Frameworks

This git repository explains how to use selected data science frameworks.

<https://github.com/data-science-and-big-data-analytics/data-science-frameworks>

A README explains how to use it. Furthermore, helpful tips and available infrastructure are stated (bwCloud, bwUniCluster, and Google Colab).

We have also added a slide deck explaining the frameworks a bit and how to use the cloud compute services available to you. Slides explaining this code (with comment function available):

<https://docs.google.com/presentation/d/1v41r4zBfYMe7okcziThfDqt0vqsKrP>

C.5 Examples of Peer-reviewed Publications from Student Submissions

Some selected publications from student submissions. Will be updated and completed shortly.

- MSc Thesis Fabian Singhofer [DocEng ‘21] (B ranked), **Best paper award!**, <https://arxiv.org/abs/2105.08842>
- Project STEREO [iiWAS’ 21] (C ranked), <https://arxiv.org/abs/2103.14124>
- Project Text Summarization [iiWAS’ 21] (C ranked), <https://arxiv.org/abs/2105.11908>
- MSc Thesis Ishwar Venugopal [IJCNN ‘21] (A ranked), <https://arxiv.org/abs/2102.07838>
- MSc Thesis Morten Jessen [DocEng ‘19] (B ranked), **Best student paper award!**, <https://dl.acm.org/doi/10.1145/3342558.3345396>
- MSc Thesis Florian Mai [JCDL ‘18] (A* ranked), <https://arxiv.org/abs/1801.06717>
- Project Quadflor: [KCAP ‘17] (A ranked), <https://arxiv.org/abs/1705.05311>
- MSc Thesis Gregor Große-Bölting [KCAP ‘15, [2]]: **Best student paper nomination!**, <https://dl.acm.org/doi/10.1145/2815833.2815838>

D EMNLP 2021 SUBMISSION GUIDELINES

FROM EMMNLP Submission Call, <https://2021.emnlp.org/call-for-papers> =====

Ethics / Impact Statement ----- Tick below if your submission contains an ethics consideration / impact statement. Note that the impact statement is optional. I/We have included an ethics / impact statement as part of our conference submission and understand that this will be taken into consideration during the review process.

Reproducibility Checklist ----- Before you submit, please make sure that the following reproducibility checklist is filled.

For all reported experimental results: -----
 A clear description of the mathematical setting, algorithm, and/or model (*) Submission of a zip file containing source code, with specification of all dependencies, including external libraries, or a link to such resources (while still anonymized) (*) Description of computing infrastructure used (*) The average runtime for each model or algorithm (e.g., training, inference, etc.), or estimated energy cost (*) Number of parameters in each model (*) Corresponding validation performance for each reported test result (*) Explanation of evaluation metrics used, with links to code (*)

For all experiments with hyperparameter search: -----
 ----- The exact number of training and evaluation runs (*) Bounds for each hyperparameter (*) Hyperparameter configurations for best-performing models (*) Number of hyperparameter search trials (*) The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy) (*) Summary statistics of the results (e.g., mean, variance, error bars, etc.) (*)

For all datasets used: ----- Relevant details such as languages, and number of examples and label distributions (*) Details of train/validation/test splits (*) Explanation of any data that were excluded, and all pre-processing steps (*) A zip file containing data or link to a downloadable version of the data (*) For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control (*)

If the above items are not applicable or if you have any additional comments, please provide your feedback below.

Note: This list is based on Dodge et al, 2019 and Joelle Pineau's reproducibility checklist. Dodge: <https://www.aclweb.org/anthology/D19-1224.pdf> Pinaue <https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf>

Further checklists for papers:

CoLLAs 2024, <https://lifelong-ml.cc/reproducibility>

NeurIPS 2021 Paper Checklist Guidelines, <https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist>

E ADMINISTRATIVE AND OTHERS

Structure of the proposal. You may well use this template also for writing the proposal of your thesis. Please make sure to cover these topics.

- Motivation
- Problem statement (incl. assumptions!)
- Research questions (separate in mandatory / optional)
- Methods (you plan to apply and/or newly develop)
- Dataset(s) (possibly also: benchmarks)
- Related work (few, key papers only in the proposal)
- Schedule (how to use the 6 months of work; commonly we use 4 months for develop, 2 for evaluation; writing starts on day 1)

Proposal is typically short, few pages (e. g., 1-2 A4 pages) in this template.

Forms for registering a thesis at UULM. MSc Thesis: https://www.uni-ulm.de/fileadmin/website_uni_ulm/studium/Studienorganisation/Pruefungsanmeldung/Formulare/antrag_masterarbeit_WEB.pdf

BSc Thesis: https://www.uni-ulm.de/fileadmin/website_uni_ulm/studium/Studienorganisation/Pruefungsanmeldung/Formulare/antrag_bachelorarbeit_WEB.pdf

And do not forget to have your signature on the paper regarding the statement of originality, see following page.

FUN

See also paper templates, but in other disciplines.

[tinyurl.com/paper-template](https://drive.google.com/file/d/1IaQpS5blxHNIKEBoXh0kQPRGjAtXr6XZ/view) → <https://drive.google.com/file/d/1IaQpS5blxHNIKEBoXh0kQPRGjAtXr6XZ/view>
and

[tinyurl.com/papertemplate](https://www.kidzone.ws/magic/walkthrough-t.htm) → <https://www.kidzone.ws/magic/walkthrough-t.htm>

Name: Space Lazer

Student Number: 666

Statement of Originality

I hereby declare that I have written the thesis by myself, without contributions from any sources or aids other than those indicated. I confirm that this work has not been submitted or published elsewhere in any other form for the fulfillment of any other degree or qualification.

.....
Place and Date

.....
Space Lazer

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009