DSBDA Template: The Name of the Title is Hope https://tinyurl.com/dsbda-template

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

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

Lars Thørväld
The Thørväld 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 Thørväld 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://anonoymo.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://anonoymo.us/me*

For the abstract, please follow the Jennifer Widom structure.

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/

*Both authors contributed equally to this research.

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/svwvwvqxfxtp 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/qxrdtnzrrpwc

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/

1.1 Motivation

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

What is the motivation?

Apply Jennifer Widom structure

1.2 Problem Statement (or: Problem Formalization)

What is the problem?

Why is it a problem?

Why is it not yet solved?

1.3 Contribution

What is our contribution?

Point of Discussion: 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.

1.4 Organization

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.

2 RELATED WORK

Use [1] or Abril and Plant [1].

But always put a tilde (~) before the \cite.

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 apparaturs given the structure below.

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

4.1 Datasets

Datasets: Which datasets do you use? Provide descriptive statistics, usually in tabluar 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?

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 the 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 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] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. https://doi.org/10.1145/1188913.1188915

[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 Manuél Gómez-Pérez (Eds.). ACM, 8:1–8:8. https://doi.org/10.1145/2815833.2815838

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 DATA SCIENCE AND BIG DATA ANALYTICS (DSBDA) GROUP

B.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

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

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

B.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_UM8rWlojgGTV11O85Lu104e2KiCBDg-k-9A

B.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.

B.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/1v41r4zBfYMe7okcziThfDqtqqsIKyPRVjNDIRQHZksRI

B.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

C 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

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

D 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

and

tinyurl.com/papertemplate → 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