DSBDA Paper 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 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

A handbook is born!

The writing template now comes with a first draft of a writing template. See `dsbda-handbook.tex` for this. The handbook has the goal to explain the template, and provide further context and guidance in writing, while not being redundant with classical scientific writing literature. It also has an extensive list of resources to books, surveys, etc.
Use the handbook as reference when writing your paper.

Abstract: How to write it

An abstract conveys in a summary of 150 words your research idea, experimental results, and their impact. It is an opportunity to directly communicate the key message of your proposal, which otherwise has to be collected from different places in the paper. With order words: *Not including an abstract in a proposal is a missed opportunity!*

This template is for papers, research-based group work reports, BSc and MSc theses, 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. For the author's information, create an ORCID and add it to your record, see the example of the first author. You can obtain an ORCID here: https://orcid.org/

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*

Note on the Use of Generative AI Tools

We are following the procedure of the German Research Foundation regarding the use of generative AI tools.

- Please carefully read the DFG's "Guidelines for Dealing with Generative Models for Text and Image Creation", which are available here: www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/stellungnahmen_papiere/2023/230921_statement_executive_committee_ki_ai.pdf
- A very good "Artificial intelligence guidance" of what one can do and what not is also found here: https://www.essex.ac.uk/student/exams-and-coursework/artificial-intelligence
- This coincides with recent regulations at international conferences such as 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.

CCS Concepts

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

Keywords

datasets, neural networks, gaze detection, text tagging

For the abstract, please follow the Jennifer Widom structure.

*Both authors contributed equally to this research.

1 Introduction

What is Strong and Ego-less Research?

Define good research questions and run experiments that generate scientific insights, i. e., new knowledge. Do not aim to develop a new method and compare it to weak baselines, cherry-picked datasets, and experimental conditions that favor your model.
Think about:

- **Baselines**: Are they strong, are they state-of-the-art?
- **Datasets**: Are they representative / used in the community, are they recent, are they challenging?
- **Related Work**: Conduct a thorough research for specific papers on the specific problem addressed by the paper. It is easy to overlook papers and with that baselines, datasets, etc.[a]
- **Tasks**: Do not consider one task only, but multiple tasks. For example, in NLP not only classification but also entity recognition; in Graph Representation Learning, not only vertex classification, but also graph classification/regression and link prediction.

[a]A statement on "Missing important related works" basically means that "the comparison with related work could be more in-depth" and that "the paper overlooks some key related works in this area".
So the related work needs to be both, covering relevant fields but also be specific to the problem. A statement like this reflects it "The related work contains broad descriptions of prior methods on [..]. It could be more focused on prior work relevant to the paper, like work involving the [...]".
In case of doubt, write more a more detailed related work and locate it in the appendix or supplementary material, respectively. A comment received was "The supplementary material is extensive and includes: Detailed proofs [...], Implementation details [...], Additional experiments [...], A comprehensive literature review and extended discussions on related works in [...]".
This makes a strong case, but only if the paper is already self-contained and the appendix is used to support the paper's claims and results. Note, a reviewer is not required to consider the appendix.

Have a throughline in your paper and maintain it!

A paper must be **consistent and coherent** in what it wants to convey to the reader. This means that you need to define and maintain a throughline in your paper.
Key place in the paper to check for coherence and consistency are

- **Title** → does it contain the key message, which is then picked up in the abstract and elaborated in the introduction,
- **Abstract**,
- **Introduction** → contributions list and research questions, respectively,
- **Datasets** → are suitable to answer the research questions from the list in the introduction,
- **Procedure** → explains the steps of the experiments taken to answer the research questions, one at a time.

Whenever you make changes at one place, check and update the others, too!

Instructions: Write following this structure.

To organize the introduction, the proposed structure of Jennifer Widom should be used. Not using the structure may leave an introduction oftentimes meaningless, when it ends at the motivation and does not well explain the *why is it a problem* and *why is it not solved* parts. 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.

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/

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?

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 are your contributions?

Instruction: Write down your list of contributions.

The introduction (and the structure of it) needs to match the bullet items of contributions at the end of the introduction. There is a clear disconnection and break in the paper if the introduction describes the motivation well, but the contributions list is about something else, see also comment below.
Your contributions list is a main point of discussion. It has to be done well.

Below, we summarize our contributions.

- 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.

Introduction What is a contribution item and what not.

The bullet items of contributions need to be a precise description of research questions that are phrased as how they make a contribution beyond the state of the art. For example, "We compare our method X with three strong baselines A, B, and C to demonstrate the effectiveness of our approach on nine benchmark datasets. [...]." The contributions list may not be a description of implementation steps, e.g., we first pre-process data, we train the models, and we evaluate the models, etc.

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.

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 always perform 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 unsual 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.),
- 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),
  or
- it is not mentioned if the procedure applied considers training a (graph) neural network in an inductive versus transductive setting (transductive models are inherently performing better on graph tasks) .

IMPORTANT: See also, and tead the summary of dozens of practices in machine learning that may invalidate the results of a research paper. "Questionable practices in machine learning", https://arxiv.org/abs/2407.12220

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 suspicous 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 <MyMethod> *or* 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

Assumptions: What are assumptions?

The assumptions describe explicitly what characteristics of the dataset, method, etc. are assumed when running the experiments. What assumptions you make are as different as the research questions. An example of an assumption in graph learning is "We assume to have access to unlabeled test nodes during training, i.e., we assume a transductive graph learning setting."

- What are the assumptions that you make?
Note: make sure there is an explicit section or subsection called "Assumptions" in your paper.

Example: A textbook example of what an assumption is

Our primary assumption [for bibliographic metadata extraction] is that all necessary information can be found within a one-hop crawl of the landing page associated with the DOI. This assumption is based on our observation that publishers present key bibliographic information on the landing page or pages directly linked to it e. g., the PDF of the publication.

Assumptions: Difference to research questions.

The assumptions are clearly not the same as the research questions (that are to be stated in the introduction). *Writing the research questions in the section on assumptions is not possible.*

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 **??**.

4.1 Datasets

Dataset: What needs to be included in the description?

The used datasets need to be described including a table showing relevant descriptive statistics. This includes the number of samples in the data set and the split of the dataset into the train, validation, and evaluation sets. Other information relevant to the experiment needs to be included such as the total number of classes and the average number of classes per sample (in case of multi-label classification), the average length of a document, etc. Commonly this information is provided in tabular form. What information is to be included depends on the research question. A good guide is to look it up from closely related papers. *Independent of what is reported on the datasets, it is always necessary to add for each average also the standard deviation.*

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

Describe the steps that are needed to prepare the datasets for the experiments. It is commonly about rather technical steps that are important for a good reproducibility of the work.

4.3 Procedure

Procedure: What needs to be described to understand the experiments.

The experimental procedure needs to be clearly described such that one can understand precisely which experiments are carried out and how. Do not mix in pre-processing (it is its own subsection above) nor implementation details (it is a subsection below). Focus on describing how the experiments are used to answer your research questions. So if there are three research questions in the order A, B, and C, one would expect that the procedure describes experiments corresponding to these research questions in exactly this order. If not already clear from the dataset description, include a clear statement about the dataset split including a rationale why this specific split is used. It can be as short as "We use a standard train/validate/test-split of 80, 10, and 10 percent of the dataset, following the literature (cite the papers)."

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?

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*" [2].

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

Ethical Statement

Write about ELSI, i. e., Ethical, Legal, and Social Implications of your research.

Instructions: How to write an ELSI statement?

If you have no idea what to write here, consult your favorite AI. Ask it for a checklist for ELSI considerations. Should you ask the AI? Is it sufficient to ask the AI?

Acknowledgments

Add this mandatory acknowledgment if you use the bwHPC.

The authors acknowledge support from the state of Baden-Württemberg through bwHPC.

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.

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] John K. Ousterhout. 2018. Always measure one level deeper. *Commun. ACM* 61, 7 (2018), 74–83. https://doi.org/10.1145/3213770

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

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 Ablation Studies

A.5 Detailed Discussions

A.6 ...

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

B Include a Checklist

New

For example, the "ACL 2023 Responsible NLP Checklist" or the reproducibility criteria of NeurIPS. Every submission must also have a section on Ethical considerations and Limitations.

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