Measuring Revision Behaviour via Similarity Metrics in Educational Settings

Andrea Horbach[1,2], Ronja Laarmann-Quante[3], Thorben Jansen[2], Johanna Fleckenstein[4],
[1] Kiel University, Germany
[2] Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3] Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Germany
[4] Hildesheim University, Germany
horbach@leibniz-ipn.de

Abstract

Revision is a core component of the writing process, especially in educational settings. This paper investigates the extent to which different textual similarity metrics between drafts and revisions can serve as indicators of revision behaviour. We evaluate a number of metrics across two distinct datasets from the German educational system: FORMAT (EFL emails) and DARIUS (argumentative essays in German). Our results show that correlations between metrics and quality improvement vary substantially between datasets. To foster more practical research, we provide a user-friendly Streamlit app that enables nontechnical users to upload text pairs and receive similarity scores.

1 Introduction

Writing is a process (Flower and Hayes, 1981), and thus revising a written draft, potentially based on feedback, is an integral part of the writing process (Fitzgerald and Markham, 1987; Liu et al., 2014) with feedback engagement being crucial to successful learning (Azevedo, 2015; Sinatra et al., 2015; Price et al., 2011; Winstone et al., 2017). Currently, there are different ways how this is operationalized, for example through time on task (Järvelä et al., 2008; Zhu et al., 2020; Bråten et al., 2022), keylogs (Leijten and Van Waes, 2013) – allowing maximal control of the writing process, but being not always available – as well as similarity or distance metrics between draft and revision (Holcomb and Buell, 2018; Cromp and Litman, 2021).

In our study, we focus on the third part, similarity measures, as these are the only measures that are available when no specific precautions have been taken to log the writing process itself. We compare similarity metrics on two (in several dimensions quite different) datasets investigating to what extent the metrics are intercorrelated and how they interact with overall text quality improvement.

From a practitioner's perspective, low threshold tools are not readily available. This is the practical gap we aim to fill. Therefore, we integrate these measures into a Streamlit app[1] allowing a non-technical user to upload texts and their revisions and download similarity scores as a CSV file. Our code is publicly available at https://github.com/andreahorbach/RevisionMeasuresStreamlitApp.

2 Related Work

In the area of argumentative writing in particular, Zhang et al. (2016) present a system that allows teachers to monitor the revision process of argumentative essays and that classifies revisions as centered on content or surface changes. Also in the context of argumentative essays, Afrin and Litman (2018) trained classifiers to decide which sentence revision is better, employing as features also similarity metrics used in our study. More recently, revision quality has also been approached using generative LLMs and, for example, chain-of-thought prompting (Liu et al., 2023).

Similar to our experiments to analyze correlations between the magnitude of revisions and grade changes, Cromp and Litman (2021) show that tf-idf-based similarity metrics between drafts and revisions capture changes in the text that also indicate grade changes. Also variants of Levenshtein distance have been used in similar contexts (Schiller et al., 2024; Miroyan et al., 2025). Similarity metrics in general have been used for a wider variety of applications, such as measuring revision behaviour in online encyclopedias (Daxenberger and Gurevych, 2013) or textual resue in general (Bär et al., 2012).

[1] https://streamlit.io

3 Experimental Setup

We use a variety of textual similarity measures to quantify revision behaviour. We refer to the first text as *draft* and the second text as *revision*.

The used measurements are:

- **Length Difference:** length of the revision minus length of the draft. Higher absolute values indicate more changes.
- **Levenshtein Distance:** the minimum number of insertions, deletions and substitutions (Levenshtein, 1966) necessary to convert the draft into the revision. Higher values indicate more changes.
- **Longest Common Substring (LCS)**: The length of the longest string that exists both in the draft and the revision. Lower LCS values indicate more changes across different locations in the text.
- **Greedy String Tiling (GST)** (Wise, 1993) finds all common substrings between draft and revision of minimum length 3. The lower the GST, the more changes have been made.[2]
- **Semantic Similarity** as measured by the cosine similarity between the two texts encoded as SBERT vectors (Reimers and Gurevych, 2019, 2020).[3] A lower value indicates more semantic changes to the text.

Length difference is measured in our app both on character and token level. Levenshtein distance, LCS and GST are measured on character, token and part-of-speech (POS) level. The POS level abstracts from the exact content and reflects structural changes. For example, if a noun is substituted with another noun in the revised text, this is a lexical change affecting the content of the text but not its syntactical structure. We normalize each measure (except for SBERT which does not have to be normalized) by the length of the draft.

4 Datasets

Out of the variety of existing revision datasets (Lee et al., 2015; Pilan et al., 2020; Kashefi et al., 2022; Velentzas et al., 2024), we conduct our experiments on two different authentic public datasets collected in schools in the German educational system. One of them contains German texts, i.e. typically the students' L1 and the other one English texts, i.e. typically the L2. As Breuer (2017) showed, revision processes can differ between L1 and L2 writing.

	FORMAT	**DARIUS Verkehr**	**DARIUS Energie**
# revisions	1,139	902	880
#tokens/text	72.7/88.4	146.1/213.8	153.4/215.5
language	L2 English	L1 German	L1 German
grade	7-9	9-13	9-13
genre	e-mails	argumentative essays	

Table 1: Overview of the datasets used and their key properties. The average number of tokens per text is given separately for drafts and revisions.

4.1 FORMAT

The FORMAT dataset (Schiller et al., 2024) contains semi-formal e-mail texts from EFL classes using the prompts described in Keller et al. (2023). Students from grade 7 to 9 were asked to first draft an e-mail asking for specific information, such as requesting information about a certain language school in the UK, and then revised it based on either automated binary feedback (see Horbach et al., 2022) about which out of 5 criteria the text already fulfilled (feedback condition) or based on a generic feedback (control condition). The dataset contains a total of 1,139 e-mail text pairs, excluding pairs where either both or one of the texts was empty. As a proxy for holistic text quality, we use the sum of 4 out of the 5 binary scores, excluding one that only pertains to the e-mail's subject line.

4.2 DARIUS

The DARIUS dataset (Schaller et al., 2024) contains 4,589 argumentative essays written by German high school students about socio-scientific issues such as energy efficiency. Each student wrote three texts. They were randomly assigned one out of two prompts (henceforth referred to as *Verkehr* 'traffic' and *Energie* 'energy') for the first writing task. Afterwards, they received feedback (either generic feedback or automated feedback based on keywords and a number of text features) and were encouraged to revise their text. Finally, they worked on the other prompt as a transfer task. Only the drafts and revisions from the first task are used by us. This leaves us with a total of 1,782 essay pairs with two non-empty texts. For holistic

[2]We use the Python package gst-calculation v.0.1.2 https://pypi.org/project/gst-calculation/.

[3]We use the model paraphrase-multilingual-MiniLM-L12-v2.

text quality, we use ability scores derived by Jansen et al. (2025) through a many-facet Rasch model.

4.3 Dataset Comparison

Table 1 shows an overview of the datasets. The FORMAT texts are considerably shorter than the DARIUS texts (see Figure 3 in Appendix A for more details on the distribution of text lengths).

5 Experiments

In the following, we evaluate the behaviour of the different revision metrics on two datasets, addressing the following research questions: How closely correlated are the individual measures on our data? How computationally expensive are they? How do they correlate with changes in text quality? The answers to these questions might help practitioners select a suitable measure for a given task without spending too much computational resources.

5.1 Analysis 1 - Correlations between revision measures

In Figure 1 we report Spearman correlations between the different measures per dataset.[4] For all measures, we see that the different variants (token, character or POS-level) are almost perfectly correlated with each other. Only for GST based on POS vs. characters the correlation is $r_s < .9$ in the DARIUS datasets.

In FORMAT, all measures are highly intercorrelated (all $r_s \geq .73$). In contrast, in both DARIUS datasets, only Levenshtein distance and length difference are highly correlated (all $r_s \geq .89$) as well as GST and LCS (all $r_s \geq .64$). The highest correlation between SBERT and any other measure in the DARIUS datasets is $r_s = .54$ (with LCS on token basis). Between GST/LCS and length difference, there is no significant correlation and between GST/LCS and Levenshtein distance, the highest correlation for DARIUS is $r_s = -.29$.

This means that it is highly dataset-dependent to what extent the different revision metrics are intercorrelated. Figure 6 in Appendix A shows the distribution of the revision measures in each dataset and we see, for example, that in FORMAT, the length difference and Levenshtein distance are much lower and have a lower variance than in DARIUS but there is a higher variance in the SBERT measure. Thus, while there are fewer changes in

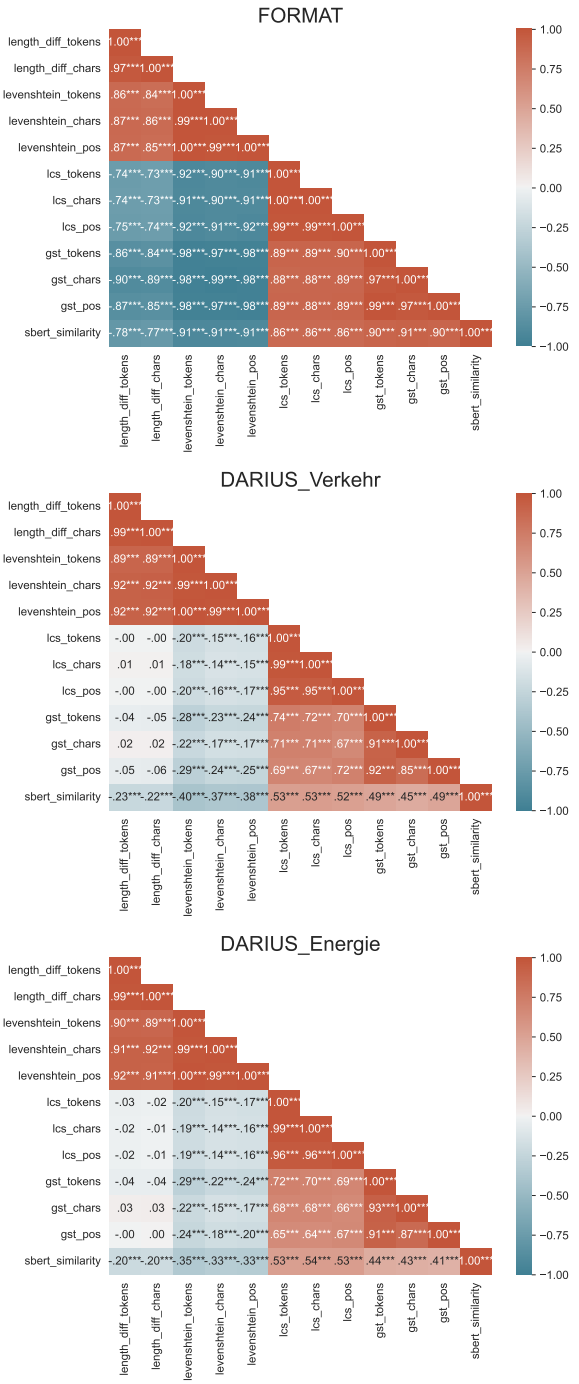

Figure 1: Spearman correlations between revision measures. Asterisks indicate significance at the 5%, 1% and 0.1% level, respectively.

[4]Note that Spearman reflects the non-linear relationship that exists between many of the measures. For comparison, the Pearson coefficients can be found in Figure 5 in Appendix A.

	FORMAT	DARIUS
length diff - chars	0 .02	0 .05
length diff - tokens	130 .30	142 .97
levenshtein - chars	11.53	126 .06
levenshtein - tokens	132 .93	141 .56
levenshtein - pos	134 .67	140 .85
lcs - chars	9.77	128 .55
lcs - tokens	136 .13	141 .13
lcs - pos	131 .33	142 .15
gst - chars	5.79	77 .88
gst - tokens	126 .71	136 .92
gst - pos	138 .70	143 .14
sbert similarity	219 .83	235 .32

Table 2: Runtime in seconds per measure and dataset for 100 pairs of draft and revision.

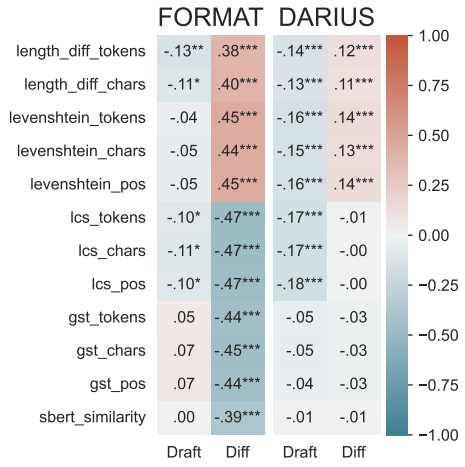

Figure 2: Spearman correlations of holistic scores with our revision measures for the draft and for the score difference between draft and revision.

the FORMAT texts overall, they tend to affect the semantics more than the changes in DARIUS.

5.2 Analysis 2 - Runtime efficiency

To give the user an idea about the resource demands of the individual measures, we performed run-time measurements on a MacBook Pro equipped with an Apple M3 chip and 16 GB of RAM. We report time to run experiments on 100 text pairs each from the DARIUS and FORMAT datasets, averaged over 10 runs. We report the average runtime in seconds in a single-threaded environment (see Table 2) per measure including the necessary preprocessing. While measures on the character level are usually more costly than those on token sequences, this advantage is often leveled out by the need for linguistic preprocessing, especially on the shorter FORMAT texts. Practitioners might use these indicators to make an informed choice which measures to compute, especially when metrics are highly correlated but vary in the computational costs.

5.3 Analysis 3 - Revisions and holistic scores

Figure 2 shows the Spearman correlation between each revision measure and a) the holistic score of the draft and b) the absolute score difference between the revision and the draft for both DARIUS prompts taken together and for FORMAT.[5] Correlations between the score of the draft and each revision measure are very low for both datasets, which means that it does not depend on the initial score how much the students will change in their text. Regarding the correlation between the score difference and the revision measures, the correlation is significant for all revision measures for FORMAT, indicating that the more revisions there are (on all levels), the more does the score change (for better or worse). For DARIUS, however, only length difference and Levenshtein distance correlate with a difference in score. Note that on average, for both datasets the revisions did not improve the scores: the median of the score difference is (close to) 0 and there are also revisions that worsened the score, see Figure 4 in Appendix A for details. Again, we see that it is extremely dependent on the dataset which measure predicts an impact of the revision process on the final score.

[5]Due to missing scores for at least one of the texts, 211 text pairs from DARIUS and 645 from FORMAT were excluded from this analysis.

6 Conclusion and Future Work

Our experiments have shown that there is not one best similarity measure to quantify revision behaviour and that the choice should be heavily influenced by the dataset at hand and the computational resources. Therefore it is important that practitioners can explore such methods freely as implemented by our Streamlit app (see Appendix B). Our goal in future work is to combine keylogging and revision measures so that changes can be tracked on a more fine-grained level (cf. Schaller et al., 2026a,b).

Limitations

The similarity measures behave differently for different languages. We already cover German and English, but logographic languages (like Chinese) or languages with a much richer morphology (such as Turkish or Finnish) might show substantially different behaviour or would require additional pro-

cessing steps such as lemmatization.

Our analysis was based on only two datasets, which showed a different behaviour with regard to the revision measures, e.g. their intercorrelations and correlations with holisitc scores. More datasets would be needed to investigate this issue more fully.

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirtenth Building Educational Applications Workshop*, page 240.

Roger Azevedo. 2015. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational psychologist*, 50(1):84–94.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING 2012*, pages 167–184, Mumbai, India. The COLING 2012 Organizing Committee.

Ivar Bråten, Natalia Latini, and Ymkje E Haverkamp. 2022. Predictors and outcomes of behavioral engagement in the context of text comprehension: When quantity means quality. *Reading and Writing*, 35(3):687–711.

Esther Odilia Breuer. 2017. Revision Processes in First Language and Foreign Language Writing: Differences and Similarities in the Success of Revision Processes. *Journal of Academic Writing*, pages 27–42.

Sonia Cromp and Diane Litman. 2021. Essay revision and corresponding grade change as captured by text similarity and revision purposes. In *Workshop for Undergraduates in Educational Data Mining and Learning Engineering*.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589.

Jill Fitzgerald and Lynda R Markham. 1987. Teaching children about revision in writing. *Cognition and instruction*, 4(1):3–24.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387.

Chris Holcomb and Duncan A. Buell. 2018. First-Year Composition as "Big Data": Towards Examining Student Revisions at Scale. *Computers and Composition*, 48:49–66.

Andrea Horbach, Ronja Laarmann-Quante, Lucas Liebenow, Thorben Jansen, Stefan Keller, Jennifer Meyer, Torsten Zesch, and Johanna Fleckenstein. 2022. Bringing automatic scoring into the classroom–measuring the impact of automated analytic feedback on student writing performance. In *Swedish Language Technology Conference and NLP4CALL*, pages 72–83.

Thorben Jansen, Lars Höft, J Luca Bahr, Livia Kuklick, and Jennifer Meyer. 2025. Constructive feedback can function as a reward: Students' emotional profiles in reaction to feedback perception mediate associations with task interest. *Learning and Instruction*, 95:102030.

Sanna Järvelä, Marjaana Veermans, and Piritta Leinonen. 2008. Investigating student engagement in computer-supported inquiry: A process-oriented analysis. *Social Psychology of Education*, 11:299–322.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane J. Litman, and Rebecca Hwa. 2022. Argrewrite v.2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, 56:881 – 915.

Stefan D. Keller, Ruth Trüb, Emily Raubach, Jennifer Meyer, Thorben Jansen, and Johanna Fleckenstein. 2023. Designing and validating an assessment rubric for writing emails in English as a foreign language. *Research in Subject-matter Teaching and Learning (RISTAL)*, 6(1):16–48.

John Lee, Chak Yan Yeung, Amir Zeldes, Marc Reznicek, Anke Lüdeling, and Jonathan Webster. 2015. CityU corpus of essay drafts of English language learners: A corpus of textual revision in second language writing. *Language Resources and Evaluation*, 49:659–683.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Ming Liu, Rafael A Calvo, Abelardo Pardo, and Andrew Martin. 2014. Measuring and visualizing students' behavioral engagement in writing activities. *IEEE Transactions on learning technologies*, 8(2):215–224.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. Predicting the quality of revisions in argumentative writing. *arXiv preprint arXiv:2306.00667*.

Mihran Miroyan, Chancharik Mitra, Rishi Jain, Gireeja Ranade, and Narges Norouzi. 2025. Analyzing pedagogical quality and efficiency of LLM responses

with TA feedback to live student questions. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 770–776.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 332–339. European Language Resources Association (ELRA).

Margaret Price, Karen Handley, and Jill Millar. 2011. Feedback: Focusing attention on engagement. *Studies in higher education*, 36(8):879–896.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024. DARIUS: A comprehensive learner corpus for argument mining in German-language essays. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4356–4367, Torino, Italia. ELRA and ICCL.

Nils-Jonathan Schaller, Thorben Jansen, Lars Höft, Hannah Pünjer, and Andrea Horbach. 2026a. GENIUS Keylog Corpus - a German High School Student Corpus with Keystroke Logging Data. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6919–6928, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Nils-Jonathan Schaller, Daniel Mora Melanchthon, Thorben Jansen, Olaf Köller, and Andrea Horbach. 2026b. KEYSCORE — keystroke-enhanced automated essay scoring. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 347–357, San Diego, California, USA. Association for Computational Linguistics.

Ronja Schiller, Johanna Fleckenstein, Ute Mertens, Andrea Horbach, and Jennifer Meyer. 2024. Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education*, 223:105163.

Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. *Educational psychologist*, 50(1):1–13.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746.

Naomi E Winstone, Robert A Nash, Michael Parker, and James Rowntree. 2017. Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational psychologist*, 52(1):17–37.

Michael Wise. 1993. String Similarity via Greedy String Tiling and Running Karp-Rabin Matching. *Unpublished Basser Department of Computer Science Report*.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B Hashemi. 2016. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations*, pages 37–41.

Mengxiao Zhu, Ou Lydia Liu, and Hee-Sun Lee. 2020. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143:103668.

A Additional Evaluations

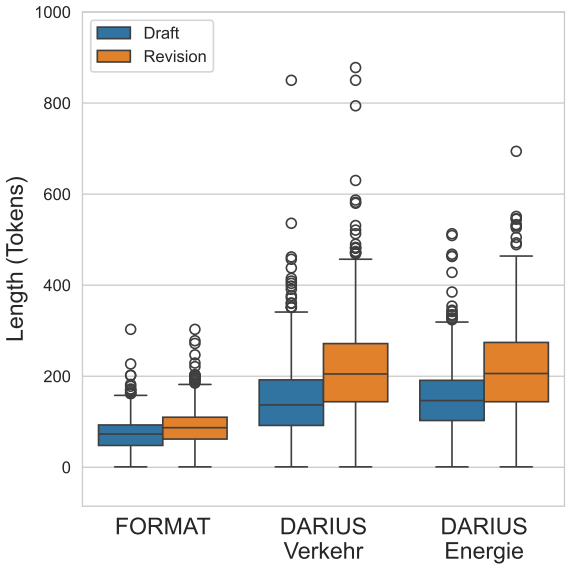

Figure 3: Distribution of text lengths (outliers > 1000 tokens cropped (3 in DARIUS_Energie, 1 in DARIUS_Verkehr).

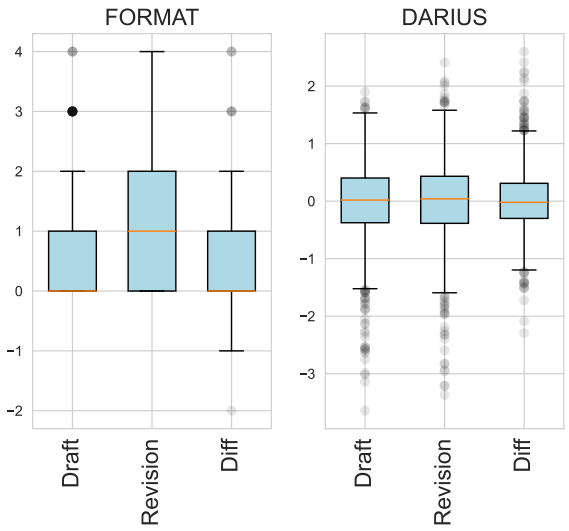

Figure 4: Distribution of holistic scores on the draft and revision, respectively, and of the score difference between draft and revision (i.e. score of the revision minus score of the draft).

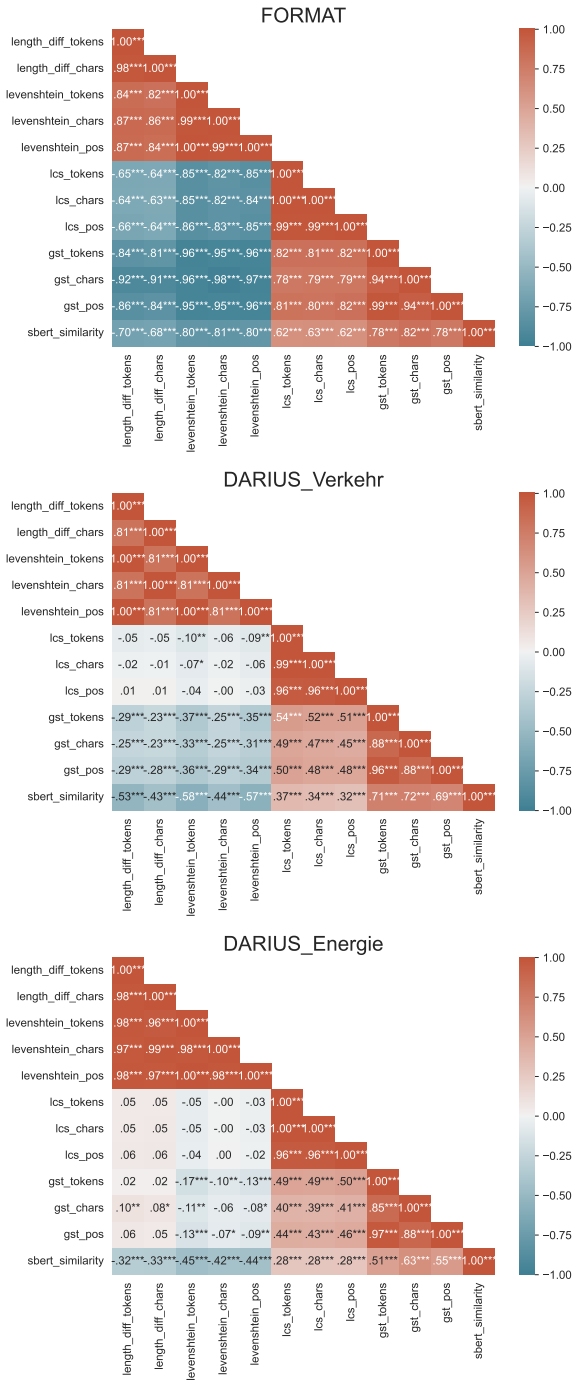

Figure 5: Pearson correlations between measures.

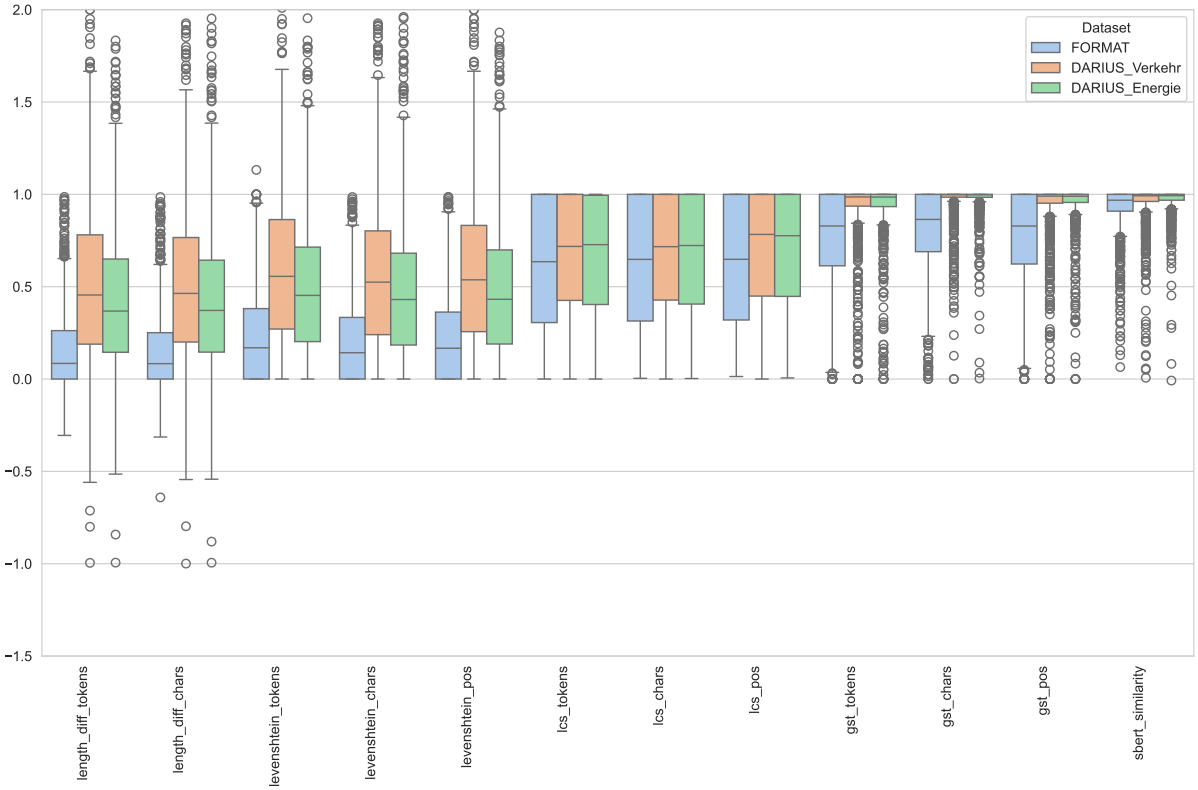

Figure 6: Distribution of revision measures per dataset. For a better visual impression, outliers ≥ 2 are cropped and displayed in Figure 7.

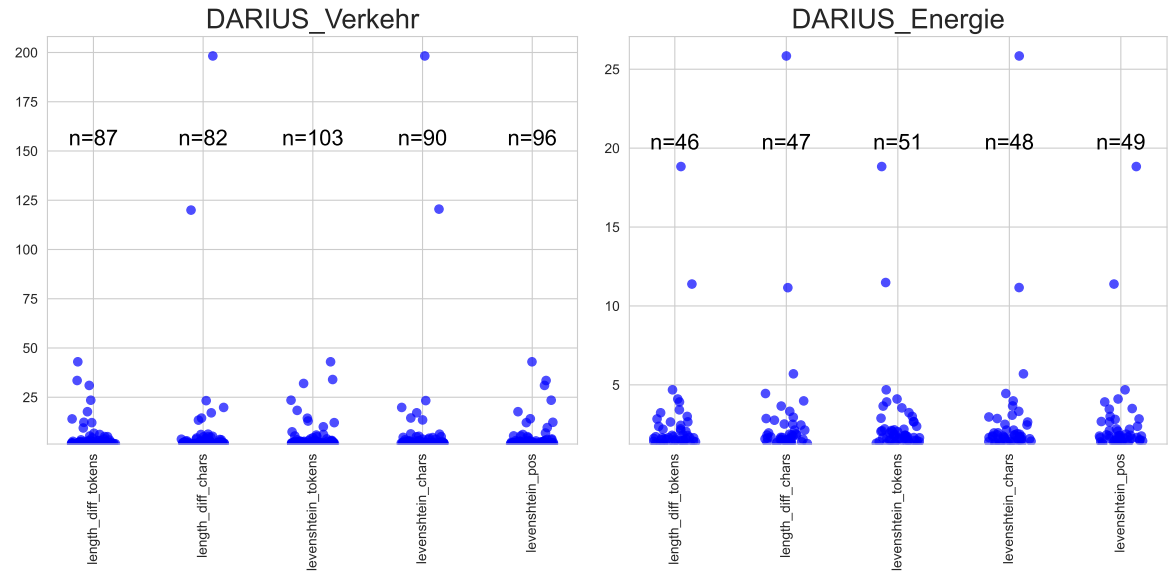

Figure 7: Outliers ≥ 2 from Figure 6.

B Streamlit App

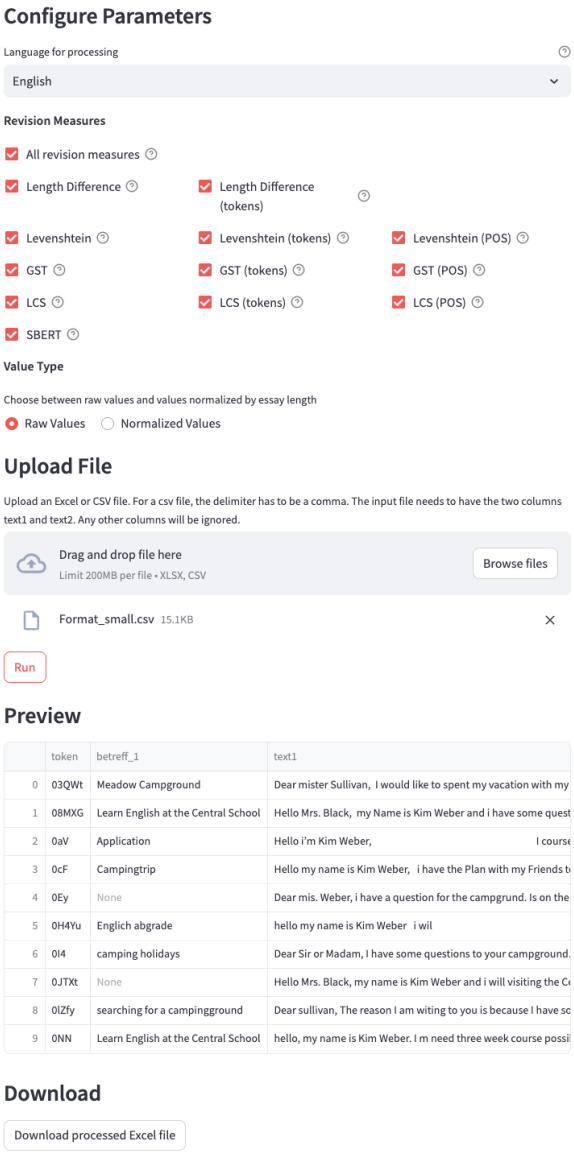

Figure 8: Sceenshot of our Streamlit app.