



Using ChatGPT-4 for the Identification of Common UX Factors within a Pool of Measurement Items from Established UX Questionnaires

CAEBUS Center for Advanced E-Business Studies

Stefan Graser (MSc) | Prof. Dr. Stephan Böhm | Dr. Martin Schrepp

14th November 2023



AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications

USER EXPERIENCE

“User’s perceptions and responses that result from the use and/or anticipated use of a system, product or service” (*DIN EN ISO 9241-210, 2020*)

- **UX is an success factor in the development and improvement of information systems** (*Rauschenberger et al. 2013, Boland 2021*)
- Multidimensional construct evaluating the overall impression (Santoso & Schrepp 2019)
- Different dimensions and quality aspects (Schrepp et al. 2023)
- „A UX quality aspect describes the subjective impression of users towards a semantically clearly described aspect of product usage or product design” (Schrepp et al. 2023)

 **Goal: creating a positive user experience** (*Boland 2021*)

MEASURING USER EXPERIENCE

- **Need to understand and measure the UX and its dimensions** to improve products, systems and services (*Irshad et al. 2020, Preece et al., 2015*)
- **Various empirical methods** can be found in literature for measuring the UX (*Preece et al. 2015, Assila et al. 2016, Albert & Tullis 2022*)
 - **Subjective methods** (self-reported data – questionnaires) or **Objective methods** (analytical data – log files)
 - Self-reported metrics most suitable to gather direct user feedback
 - Applying questionnaires: quickly, simply and cost-effectively

USER EXPERIENCE QUESTIONNAIRES

- **40 established** UX questionnaires (*Schrepp 2020*)
- Questionnaires are based on **different dimensions (factors), items, and scales** in relation to the UX (*Hinderks et al. 2019, Schrepp 2020, Schrepp et al. 2023*)
- Break down the construct UX in different factors measured by items and scales
- Measurement items characterize the user's subjective impression
- Existing questionnaires differ in the **dimensions (factors), items, and scales** (*Hinderks et al. 2019, Schrepp 2020, Schrepp et al. 2023*)

Item	Factor	Questionnaire
The system is easy to use	Likeability	SASSI
I thought the system was easy to use	Usability	SUS
[This system] is easy to use	Overall	UMUX
it was simple to use this system	System Usefulness	PSSQU

 **UX factors with different names can measure the same thing, but factors with the same name can also measure different aspects**

SEMANTIC AND EMPIRICAL SIMILARITY

- Semantic similarity refers to the degree of likeness or resemblance between the item texts based on their meaning (*Mikolov et al. 2013, Kenter et al. 2016, Conneau et al. 2018*).
- Empirical similarity refers to the degree of likeness based on measurable characteristics → thus semantically different items can refer to the same
 - **Differentiation between semantic and empirical similarity**
 - **Semantically different aspects can show a high empirical similarity**
- E.g. Items items in relation to usability correlate with items of beauty (Ilmberger et al. 2008, Tuch et al. 2012)
 - **Caused by different affects and common aspects** (Lance et al. 1994, Ford and Smith 1987, Norman 2004, Ngo et al. 2000, Bonsiepe 1968)

RESEARCH OBJECTIVE

- **Difference between the semantic similarity and empirical similarity**
- Focusing on the semantic structure of the textual measurement items
→ **Semantic Textual Similarity (STS)**
- Identifying a common ground on UX measurement item level
- **Semantic Textual Similarity (STS)** analysis using Generative AI
- ➔ **Identify semantic similar UX concepts applying GenAI**

RELATED RESEARCH

- Different Natural Language Processing (NLP) methods for Semantic Textual Similarity (STS) (Li et al. 2006, Luhn 1957, Spärck Jones 2004, Gatford 1995, Deerwester et al. 1990, Le and Mikolov 2014, Reomers and Gurevych 2019, Takhur et al. 2020, Sun et al. 2022)
- Based on encoded word embeddings in a vector space
- 3 articles aimed to consolidate UX factors based on empirical similarity (Winter et al. 2015, Hinderks et al. 2020, Schrepp et al. 2023)
- 2 NLP approaches applying Augmented SBERT and BERTopic (Topic Modeling) (Graser and Böhm 2023a , Graser and Böhm 2023b)

 **Only few research applying NLP techniques and no article using Generative AI for measuring STS**

Transformer:

A model architecture for NLP-tasks

Encoder:

A model which transforms data into a another representation format, e.g. transforming sentences into a vector in a vector space

Embedding:

Vector representation of data

RQ1: Is Generative AI able to identify useful similarity topics based on measurement items?

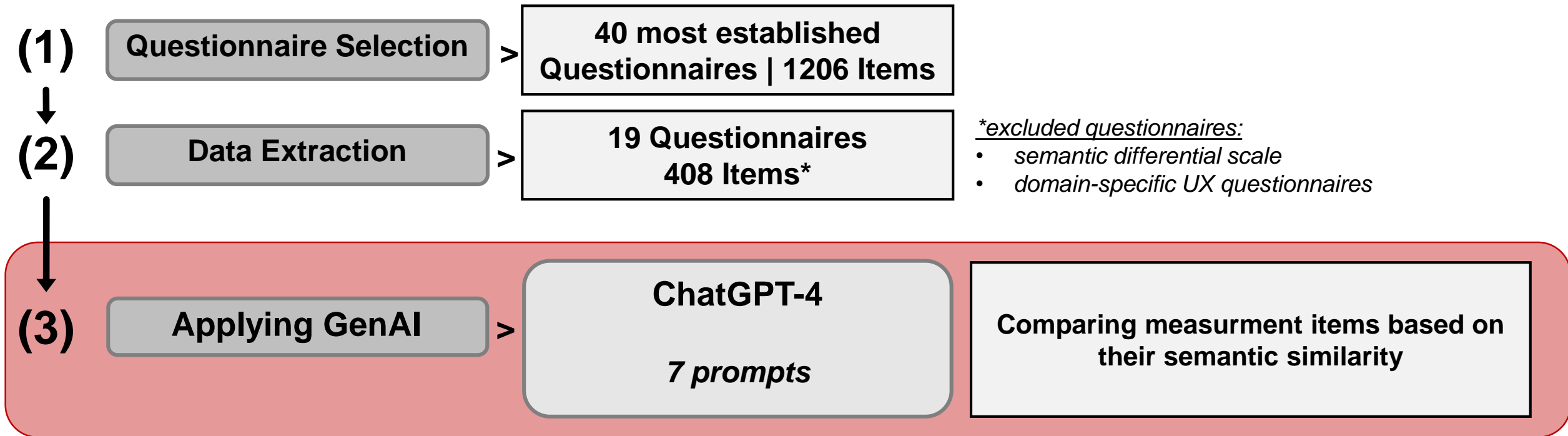
RQ2: Which topics based on semantically similar measurement items can be identified among the most established UX questionnaires?



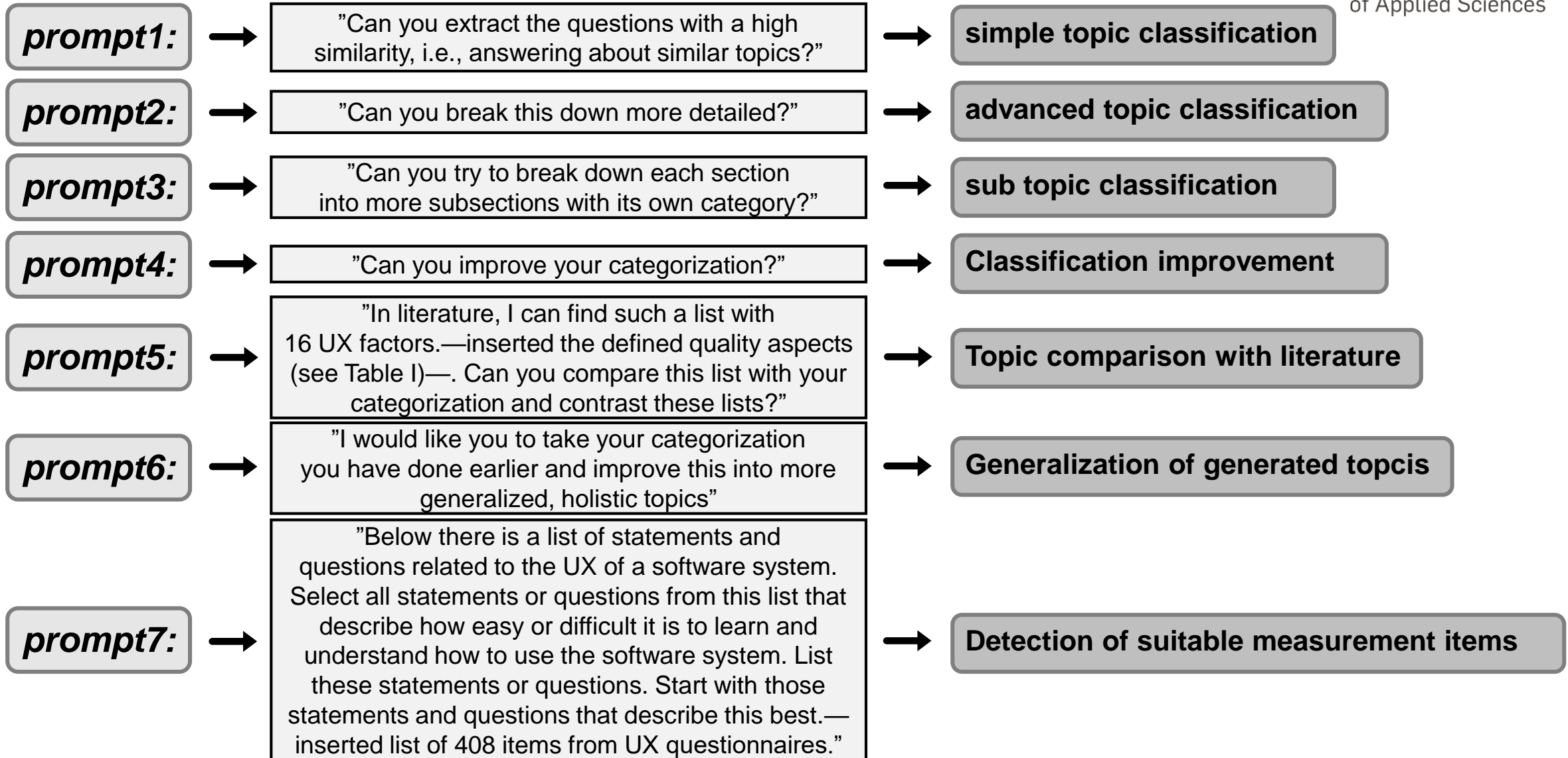
AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications

METHODOLOGICAL APPROACH



PROMPTS FOR CHATGPT-4





AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications

RESULTS – PROMPT 1

- (1) Usability and Ease of Use
- (2) Design and Aesthetics
- (3) User Engagement and Experience
- (4) Trust and Reliability
- (5) Information Access and Clarity
- (6) Issues and Errors

→ Functional as well as emotional topics were generated

→ Item classification was plausible

→ Categorizations are very broad, e.g., *Usability and Ease of Use*

 **GenAI can identify logical topics**

prompt_1:

"Can you extract the questions with a high similarity, i.e., answering about similar topics?"

RESULTS – PROMPT 2

- (1) Ease of Use
- (2) Complexity and Usability Issues
- (3) Design and Appearance
- (4) Engagement and Immersion
- (5) Performance and Responsiveness
- (6) Reliability and Trust
- (7) Information Quality and Access
- (8) Errors and Bugs
- (9) Learning and Memorability
- (10) Effectiveness and Efficiency

→ More precious classification

→ Functional, task-related topics were further broken down

Pragmatic: (1), (2), (3), (4), (7), (8), (9), (10) // **Hedonic**: (3) and (4) //

Topic (6) contains **both pragmatic and hedonic** items

 **GenAI can distinguish topics even more precisely**

prompt_2:

"Can you break this down more detailed?"

RESULTS – PROMPT 3

prompt_3:

”Can you try to break down each section into more subsections with its own category?”



RheinMain University
of Applied Sciences

- **Ease of Use:** System Usability—Website Usability—Application Usability
- **Complexity and Usability Issues:** System Complexity—Frustration and Difficulty— System Limitations
- **Design and Appearance:** Visual Attraction—Layout and Structure—Design Consistency
- **Engagement and Immersion:** Time Perception and Involvement—Depth of Experience
- **Performance and Responsiveness:** Speed of Response
- **Reliability and Trust:** Website Trustworthiness—System Reliability
- **Information Quality and Access:** Quality of Information—Accessibility of Information
- **Errors and Bugs:** Technical Issues—Error Messages
- **Learning and Memorability:** Learning Curve—Recall and Retention
- **Effectiveness and Efficiency:** Functional Efficiency—Expected Functionality

→ 10 main topics and 22 sub topics

 **GenAI can break down logical subtopics concerning the respective item focus**

RESULTS – PROMPT 4

prompt_4:
"Can you improve your categorization?"

- **System Usability and Performance:** Ease of Use—Efficiency and Speed— Functionality and Flexibility
- **User Engagement and Experience:** Engagement Level—Aesthetics and Design—Confusion and Difficulty
- **Information and Content:** Clarity and Understandability—Relevance and Utility—Consistency and Integration
- **Website-specific Feedback:** Navigation and Usability—Trust and Security— Aesthetics and Design
- **Learning and Adaptability:** Learning Curve—Adaptability
- **Overall Satisfaction and Recommendation:** Satisfaction—Recommendation

→ Emphasizes two-level structure of the main and subtopics

→ Rather broad

→ Main topic contain sub topics with pragmatic as well as hedonic properties

 **GenAI can improve topics and sub topics.**

RESULTS – PROMPT 5

prompt_5:

"In literature, I can find such a list with 16 UX factors.—inserted the defined quality aspects (see Table I)—. Can you compare this list with your categorization and contrast these lists?"

(#)	UX Quality Aspects (based on Schrepp 2023)	AI-generated (Sub) Topics
1	Perspicuity	Ease of Use—Learning Curve
2	Efficiency	Efficiency and Speed
3	Dependability	Consistency and Integration
4	Usefulness	Functionality and Flexibility—Relevance and Utility
5	Intuitive Use	Ease of Use
6	Adaptability	Adaptability
7	Novelty	-
8	Stimulation	Engagement Level
9	Clarity	Clarity and Understandability
10	Quality of Content	Relevance and Utility
11	Immersion	Engagement Level
12	Aesthetics	Aesthetics and Design—Aesthetics and Design
13	Identity	-
14	Loyalty	Loyalty
15	Trust	Trust and Security
16	Value	Perceived value

RESULTS – PROMPT 5

- UX Quality Aspects ***Novelty*** & ***Identity*** are not covered
- Overlap of Items within some AI-generated topics
- Results of the literature are more generalized (e.g. Trust)

 **Categorization of GenAI has a stronger focus on the pragmatic property and is more specific**

RESULTS – PROMPT 6

Exemplary first topic:

Usability and Intuitiveness

Ease of Use:

- 1) The system is easy to use. (+)
- 2) I think that I would need the support of a technical person to be able to use this system.(+)
- 3) I would imagine that most people would learn to use this system very quickly.(+)
- 4) I learned to use it quickly.(+)
- 5) I can use it without written instructions.(+)

Efficiency and Speed:

- 1) The interaction with the system is fast.(+)
- 2) The system responds too slowly.(+)
- 3) This software responds too slowly to inputs.(+)
- 4) The speed of this software is fast enough.(+)
- 5) has fast navigation to pages.(+)

prompt_6:

"I would like you to take your categorization you have done earlier and improve this into more generalized, holistic topics"



RheinMain University
of Applied Sciences

(+) // suitable item
(-) // unsuitable item
(+ -) // also refers to other topic
***done by researcher**

Adaptability:

- 1) The system is too inflexible.(+)
- 2) This software seems to disrupt the way I normally like to arrange my work.(+)
- 3) It is flexible.(+)
- 4) It requires the fewest steps possible to accomplish what I want to do with it.(+- Efficiency)
- 5) It is relatively easy to move from one part of a task to another.(+- Efficiency)

RESULTS – PROMPT 6

- Two-dimensional separation into the main topic and sub-topics
- Both pragmatic as well as hedonic aspects are captured
- Mostly, the items are coherent with each other and fit the construct (especially pragmatic topics) → (+ -)



ChatGPT performs very well in consolidating and developing topics concerning a holistic view of UX

RESULTS – PROMPT 7

Detected Items in relation to *Perspicuity / Learnability / Ease of Learning*:

- 1) It was easy to learn to use this system
- 2) I could effectively complete the tasks and scenarios using this system
- 3) I was able to complete the tasks and scenarios quickly using this system
- 4) I felt comfortable using this system
- 5) The system gave error messages that clearly told me how to fix problems
- 6) Whenever I made a mistake using the system, I could recover easily and quickly
- 7) The information provided with this system (online help, documentation) was clear
- 8) It was easy to find the information I needed
- 9) The information provided for the system was easy to understand
- 10) The information was effective in helping me complete the tasks and scenarios
- 11) The system was easy to use from the start
- 12) How the system is used was clear to me straight away
- 13) I could interact with the system in a way that seemed familiar to me
- 14) It was always clear to me what I had to do to use the system
- 15) The process of using the system went smoothly



RheinMain University
of Applied Sciences

prompt_7:

”Below there is a list of statements and questions related to the UX of a software system. Select all statements or questions from this list that describe how easy or difficult it is to learn and understand how to use the software system. List these statements or questions. Start with those statements and questions that describe this best.—inserted list of 408 items from UX questionnaires.”

 **GenAI can easily and precisely detect suitable items.**



AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications

DISCUSSION & LIMITATIONS

- Exclusion of common questionnaires
- Explorative deterministic → never equal results
- No adjustment of any parameters concerning ChatGPT

Implications

RQ1: Is Generative AI able to identify useful similarity topics based on measurement items?

→ GenAI can be used to (1) **classify** items from UX questionnaires concerning their semantic meaning, (2) **improve** and **compare** classifications, and (3) **detect** and **assign** items to classified topics.

RQ2: Which topics based on semantically similar measurement items can be identified among the most established UX questionnaires?

→ **6 main topics and 15 subtopic were identified**

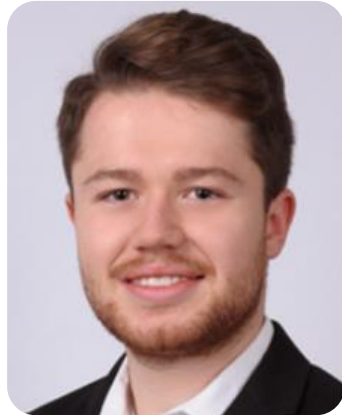
OUTLOOK AND FUTURE WORK

Further research applying GenAI and LLMs:

- Prompt engineering for further investigations
- Empirical validation of classified items in relation to the AI-generated topics
- Development of an AI-generated questionnaire and comparison towards existing UX questionnaires
- Adjustment and modification in relation to different application fields/scenarios
- AI-based item generation
- AI-based question guidance systems (instead of standardized questionnaires)

→ new way to define and break down the construct UX by differentiating between empirical and semantic similarity

THANK YOU FOR YOUR ATTENTION!



Stefan Graser (MSc)
Doctoral Candidate & Research
Associate



Prof. Dr. Stephan Böhm
Professor for Telecommunication and
Mobile Media



Dr. Martin Schrepp
UX Expert & Researcher
SAP SE



Research:

- User Experience
- Mobile Augmented Reality
- Digital Transformation / E-Business

- Mobile Media & Business / Media Innovation & Technologies
- User-Centered Design / Technology Acceptance / E-Business

- User Experience
- UX Questionnaires
- UX Research

Contact:

E-Mail: stefan.graser@hs-rm.de

Phone: +49 611 9495 2248

E-Mail: stephan.boehm@hs-rm.de

Phone: +49 611 9495 2212

E-Mail: martin.schrepp@sap.com

Connect!



REFERENCES (1/6)

1. Rauschenberger, M., Schrepp, M., Cota Pérez, M., Olschner, S., Thomaschewski, J.: Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (ueq) (2013).
2. DIN EN ISO 9241-210.: Ergonomics of human-system interaction – part 210: Human-centred design for interactive systems (2010).
3. Hassan, H. M., Galal-Edeen, G. H.: From usability to user experience. In: International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), pp. 216-222. IEEE (2017).
4. Hinderks, A., Winter, D., Schrepp, M., Thomaschewski, J.: Applicability of user experience and usability questionnaires. In: Journal of Universal Computer Science, 25 (13), pp. 1717-1735 (2019).
5. Schrepp, M.: A comparison of ux questionnaires – what is their underlying concept of user experience?. In: Mensch und Computer 2020 Workshopband (2020).
6. Santoso, H. B., Schrepp, M.: The impact of culture and product on the subjective importance of user experience aspects. Helion 5, 1-12 (2019).
7. Assila, A., Oliveira, K., Ezzedine, H.: Standardized usability questionnaires: Features and quality focus. In: Computer Science and Information Technology, 6 (1), 15-31 (2016).

REFERENCES (2/6)

8. Albert, B., Tullis, T.: Measuring the user experience: Collecting, analyzing, and presenting ux metrics. Morgan Kaufmann (2022).
9. M. Schrepp, J. Kollmorgen, A.-L. Meiners, A. Hinderks, D. Winter, H. B. Santoso, J. Thomasche-wski. On the Importance of UX Quality Aspects for Different Product Categories, International Journal of Interactive Multimedia and Artificial Intelligence, (2023).
10. Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2020). Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. *ArXiv, abs/2010.08240*.
11. C.D. Manning, P. Raghavan and H. Schütze (2008). Introduction to Information Retrieval. Cambridge University Press.
12. Reimers, N. (2022). Clustering. Online: <https://www.sbert.net/examples/applications/clustering/README.html#fast-clustering>. [Accessed: 05.2023].
13. Maaten, L.v.d., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol 9, pp. 2579—2605.
14. Wattenberg, M., Viégas, F., Johnson, I. How to Use t-SNE Effectively, *Distill*. (2016).
15. Reimers, N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Conference on Empirical Methods in Natural Language Processing*. (2019).

REFERENCES (3/6)

16. Reimers, N. (2022). Augmented SBERT. Online: https://www.sbert.net/examples/training/data_augmentation/README.html. [Accessed: 05.2023].
17. Reimers, N. (2022). Pretrained Models. Online: https://www.sbert.net/docs/pretrained_models.html#sentence-embedding-models. [Accessed: 05.2023].
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, B., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85), pp. 2825–2830.
19. Allwright, S. (2022). Which are the best clustering metrics? (explained simply), <https://stephenallwright.com/good-clustering-metrics/>. [Accessed: 06.2023].
20. Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53-65.
21. Calinski, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*.
22. Davies, D. L., Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-1 (2): 224-227
23. Grootendorst, M.R. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv, abs/2203.05794*.
24. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, retrieved: 10/2023, 2013. eprint: 1310.4546. [Online]. Available: <https://arxiv.org/abs/1310.4546>.
25. T. Kenter, A. Borisov, and M. de Rijke, Siamese cbow: Optimizing word embeddings for sentence representations, 2016. eprint: 1606.04640.
26. A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, Supervised learning of universal sentence representations from natural language inference data, 2018. eprint: 1705.02364

REFERENCES (4/6)

27. W. Ilmberger, M. Schrepp, and T. Held, "Cognitive processes causing the relationship between aesthetics and usability," in HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4, Springer, 2008, pp. 43–54.
28. A. N. Tuch, E. E. Presslauer, M. Stöcklin, K. Opwis, and J. A. Bargas-Avila, "The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments," *International journal of human-computer studies*, vol. 70, no. 11, pp. 794–811, 2012.
29. A. N. Tuch, E. E. Presslauer, M. Stöcklin, K. Opwis, and J. A. Bargas-Avila, "The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments," *International journal of human-computer studies*, vol. 70, no. 11, pp. 794–811, 2012.
30. C. E. Lance, J. A. LaPointe, and A. M. Stewart, "A test of the context dependency of three causal models of halo rater error.," *Journal of Applied Psychology*, vol. 79, no. 3, pp. 332–340, 1994.
31. G. T. Ford and R. A. Smith, "Inferential beliefs in consumer evaluations: An assessment of alternative processing strategies," *Journal of consumer research*, vol. 14, no. 3, pp. 363–371, 1987.
32. D. A. Norman, *Emotional design: Why we love (or hate) everyday things*. Civitas Books, 2004.
33. D. C. L. Ngo, L. S. Teo, and J. G. Byrne, "Formalising guidelines for the design of screen layouts," *Displays*, vol. 21, no. 1, pp. 3–15, 2000.
34. G. Bonsiepe, "A method of quantifying order in typographic design," *Visible Language*, vol. 2, no. 3, pp. 203–220, 1968.

REFERENCES (5/6)

35. Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138–1150, 2006.
36. H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.
37. K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 60, no. 5, pp. 493–502, 2004.
38. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, pp. 109–126, 1995.
39. S. Deerwester, S. T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.
40. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.
41. N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Conference on Empirical Methods in Natural Language Processing*, 2019.
42. N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks," *arXiv preprint arXiv:2010.08240*, Oct. 2020.
43. X. Sun et al., "Sentence Similarity Based on Contexts," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 573–588, 2022, ISSN: 2307-387X. DOI: 10.1162/tacl a 00477.

REFERENCES (6/6)

44. S. Graser and S. Böhm, “Quantifying user experience through self-reporting questionnaires: A systematic analysis of sentence similarity between the items of the measurement approaches,” in Lecture Notes in Computer Science, LNCS, volume 14014, Springer Nature, 2023.
45. S. Graser and S. Böhm, “Applying augmented sbert and bertopic in ux research: A sentence similarity and topic modeling approach to analyzing items from multiple questionnaires,” in Proceedings of the IWEMB 2023, Seventh International Workshop on Entrepreneurship, Electronic, and Mobile Business, 2023.
46. D. Winter, M. Schrepp, and J. Thomaschewski, “Faktoren der user experience: Systematische “übersicht “über produktrelevante ux-qualitätsaspekte,” in Workshop, A. Endmann, H. Fischer, and M. Krökel, Eds. Berlin, München, Boston: De Gruyter, 2015, pp. 33–41, ISBN: 9783110443882. DOI: doi:10.1515/9783110443882-005.