

# Task-based Evaluation of Sentiment Visualization Techniques

Kostiantyn Kucher<sup>1,2</sup><sup>a</sup>, Samir Bouchama<sup>1,3</sup><sup>b</sup>, Achim Ebert<sup>3</sup><sup>c</sup> and Andreas Kerren<sup>1,2</sup><sup>d</sup>

<sup>1</sup>Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden

<sup>2</sup>Department of Science and Technology, Linköping University, Norrköping, Sweden

<sup>3</sup>Computer Graphics and HCI Group, Technical University of Kaiserslautern, Kaiserslautern, Germany

**Keywords:** Sentiment Visualization, Sentiment Analysis, Visual Variable, Visual Representation, Visual Encoding, User Study, Text Visualization, Visual Analytics, Information Visualization.

**Abstract:** Sentiment visualization is concerned with visual representation of sentiments, emotions, opinions, and stances typically detected in textual data, for example, charts or diagrams representing negative and positive opinions in online customer reviews or Twitter discussions. Such approaches have been applied for the purposes of academic research and practical applications in the past years. But the question of usability of these various techniques still remains generally unsolved, as the existing research typically addresses individual design alternatives for a particular technique implementation only. This work focuses on evaluation of the effectiveness and efficiency of common visual representations for low-level visualization tasks in the context of sentiment visualization. More specifically, we describe a task-based within-subject user study for various tasks, carried out as an online survey and taking the task completion time and error rate into account for most questions. The study involved 50 participants, and we present and discuss their responses and free-form comments. The results provide evidence of strengths and weaknesses of particular representations and visual variables with respect to different tasks, as well as specific user preferences, in the context of sentiment visualization.


## 1 INTRODUCTION


*Information visualization* (InfoVis) is widely used as part of various data analysis and presentation workflows, with a strong interest demonstrated within and beyond the academic environment (Fekete et al., 2008). In order to reach the goals intended by the users or the audience of such visualization approaches, design choices have to be made with respect to the task, data, and limitations of human perception and cognition (Carpendale, 2003; Amar and Stasko, 2005; Munzner, 2015). All of these concerns are valid and relevant within the area of *sentiment visualization*, which is concerned with (interactive) visual representation of sentiments, emotions, opinions, and stances. While these concepts are not exactly identical (Munezero et al., 2014), they are related and are often used interchangeably as part of practical applications dealing with detection and analysis of subjectivity. The main modality of interest for such analyses


is text data, and the possible data sources and applications include customer reviews, social media posts, etc. Various computational approaches have been proposed to detect and categorize sentiment in such data. The typical task is to classify the polarity/valence of a given text item (sentence, document, etc.) as *negative*, *neutral*, or *positive*, although further alternatives exist with regard to the categories (e.g., emotions such as *anger*), scope of analysis, etc. (Mohammad, 2016). The visualization techniques complementing such computational analyses have also been discussed in the literature (Shamim et al., 2015; Kucher et al., 2018). However, the previous works on sentiment visualization have been limited with respect to the evidence on usability (Frøkjær et al., 2000) of such techniques for specific tasks. Discovering the respective evidence from empirical data would be an InfoVis *evaluation* task (Isenberg et al., 2013).


In this work<sup>1</sup>, we aim to address the research gap on sentiment visualization evaluation by collecting and analyzing the evidence of the usability of common design alternatives for particular user tasks. More specifically, we focus on

<sup>1</sup>Based on a thesis project (Bouchama, 2021).

<sup>a</sup> <https://orcid.org/0000-0002-1907-7820>

<sup>b</sup> <https://orcid.org/0000-0002-6160-3687>

<sup>c</sup> <https://orcid.org/0000-0001-7938-6732>

<sup>d</sup> <https://orcid.org/0000-0002-0519-2537>

the questions of effectiveness and efficiency of *visual metaphors/representations* and *visual variables/channels* for visual encoding of sentiment and emotions. These research questions are relevant to the basic choices affecting the design of visualization techniques. To answer them, we design and conduct a task-based within-subject user study involving visual representations of sentiment, while taking the error rate and task completion time into account. The study was carried out online, and the responses and free-form feedback were collected from 50 participants in May–June 2021. The findings of our work can be applied for choosing sentiment visualization strategies for communicating the results of computational and/or manual sentiment analysis methods.

## 2 RELATED WORK

In this section, we discuss the important concepts and previous works relevant to the problem of evaluating sentiment visualization.

### 2.1 Visual Representations

The choice of particular visual representations/metaphors is dependent on the context of their usage. As Görg et al. discuss, there are multiple different visual representations available—from simple and complex over to univariate and multivariate representations, with theoretical design considerations and empirical evidence suggesting the feasibility (and usability) of particular representations for particular tasks (Görg et al., 2007). Common visual representations are, for example, bar charts, scatter plots, pie charts, line charts, etc.

Visual representations make use of low-level graphic elements (marks) and visual variables to convey the information to the user. Visual variables represent attributes of graphical marks that are easily processed by the human (Görg et al., 2007). Bertin defines the following seven visual variables: *position*, *form*, *orientation*, *color*, *texture*, *value*, and *size* (Bertin, 2011). These variables are distinguished perceptually without the use of cognitive steps in contrast to comparing written numbers, for instance (Görg et al., 2007). However, choosing the right visual variables depends on the underlying data that is to be visualized. Munzner discusses two important principles for using visual channels and representations: *expressiveness*, meaning that the encoding should aim to represent all the information present in the data, without misleading the user; and *effectiveness*, meaning that the importance of the data attribute

should match the saliency/noticeability of the visual channel (Munzner, 2015). These considerations will play an important role in this work.

### 2.2 Sentiment Visualization Techniques

The existing sentiment visualization techniques have been discussed in the literature in the respective reviews and surveys (Boumaiza, 2015; Shamim et al., 2015; Kucher et al., 2018) as well as in the larger context of text visualization, visual text analytics, and social media visual analytics (Kucher and Keren, 2015; Chen et al., 2017; Alharbi and Laramee, 2019). The prior research establishes several categorizations of such sentiment visualization techniques, which typically include the dimension of the source data domain (such as customer reviews or social media posts), the dimension of sentiment categories or quantitative values (e.g., *positive* vs *negative*, a list of basic emotions, etc.), and the choice of visual representations used by the respective technique. Common visual representations such as bar charts and line charts can be used for these purposes, but also novel metaphors (Shamim et al., 2015). Furthermore, some of the existing surveys discuss the choice of visual variables for representing the actual sentiment categories or values. Kucher et al. report that the majority of techniques in their survey use the visual variable of color for this purpose, followed by position/orientation and size/area (Kucher et al., 2018). Whether a particular combination of the visual representation and visual variable fits the user tasks associated with sentiment visualization is then the question to answer—and in order to provide the respective guidelines not only from the position of general-purpose information visualization principles, but sentiment visualization in particular, there is a need to collect further empirical evidence, as discussed next.

### 2.3 Evaluation of InfoVis Approaches

Evaluation is typically mentioned in visualization research as an umbrella term for various forms of validation, ranging from use cases and domain expert reviews to longitudinal case studies and controlled lab experiments (Carpendale, 2008; Lam et al., 2012; Isenberg et al., 2013; Elmqvist and Yi, 2015). In human-computer interaction (Ebert et al., 2012), “evaluation” (focusing on estimating the usability and collecting user feedback for mainly formative purposes as part of an iterative design-implementation-validation process) is sometimes contrasted to “experiment” (typically summative purposes and a controlled environment) (Purchase, 2012).

Regarding examples of existing evaluations of sentiment visualization, for instance, Diakopoulos et al. discuss the feedback of the target audience (journalists) of their tool Vox Civitas focusing on the perceived effectiveness and usefulness (Diakopoulos et al., 2010). Zhao et al. complement a more open-ended participatory study with a crowdsourced task-based study for their emotion visualization approach PEARL, providing empirical data on the usability of the tool for particular tasks (Zhao et al., 2014).

Prior results covering multiple techniques and representations for sentiment visualization are limited, though. The review by Boumaiza includes multiple techniques, but does not provide a systematic categorization or comparison (Boumaiza, 2015). The work by Shamim et al. reveals findings on the usability of opinion mining systems' visualizations (Shamim et al., 2015). They classify 11 techniques according to the visual metaphor, including bar charts, rose plot variations, etc. Their work compares techniques concerning different metrics such as being eye-pleasing, easy-to-understand, user-friendly, etc. They conducted a questionnaire survey, and data was collected via this questionnaire and through seminars. They conclude that simple, easy-to-understand, low-dimensional visualizations are rated higher than the others. Shamim et al. rank the top five representations included in their study as follows, according to the users' preferences: bar chart, glowing bar, treemap, line graph, and pie chart. These results are interesting, but while they are based on the reported preferences, they do not provide evidence about the *usability* (Frøkjær et al., 2000) of such representations for user tasks in the context of sentiment visualization. In our study, we aim to focus on two aspects of usability: effectiveness and efficiency. As discussed by Frøkjær et al., *effectiveness* is assessed by determining accuracy and completeness with which users achieve certain goals; for instance, this could be achieved by measuring the error rate of user responses compared to the ground truth data available to the experimenter (Purchase, 2012). *Efficiency* is calculated by the relation between the accuracy and completeness with regard to the resources used to complete the task, for instance, the task completion time. These aspects are taken into account as we continue the discussion of particular user tasks for our study.

### 3 STUDY DESIGN

In this section, we describe the experimental design of our study, including the particular user tasks, visual representations, datasets, and implementation details.

#### 3.1 User Tasks and Alternatives

Our within-subject study was targeted at the general public to facilitate the generalizability of the findings and to increase the potential number of participants. This required all the tasks and visualizations to be intuitive to the average user, whose level of visualization literacy might be limited. The design of our study was inspired by the previous work on task-based evaluation of common visual representations (Saket et al., 2019), and it involved several user tasks related to sentiment visualization in our case. To visualize sentiment and emotion, we used several sentiment analysis datasets depending on the task. Based on these visualizations, we came up with different survey questions for each plot. The particular choice of the tasks was based on the work by Amar et al., who proposed a set of ten low-level analysis tasks that describe users' activities while using visualization tools to understand their data (Amar and Stasko, 2005). In our study, we focused on a subset of seven tasks relevant to sentiment visualization and common visual encodings:

- *Find Anomalies*: e.g., *Does the data look abnormal?*
- *Find Clusters*: e.g., *How many clusters can you identify in this plot?*
- *Find Correlation*: e.g., *Is there a (weak/strong) correlation between sentiment intensity and time?*
- *Compute a Derived Value*: e.g., *What is the sum of all negative and neutral sentiments in the pie chart?*
- *Find Extremum*: e.g., *What sentiment has the highest/lowest value?*
- *Filter*: e.g., *How many emotional states can you identify between time step X and Y?*
- *Retrieve Value*: e.g., *Which emotion has green color in this plot?*

To address the research questions of our work, we aimed to propose the study participants to solve these tasks while using various combinations of visual representations and visual variables. In particular, we focused on the following five standard visual representations: scatter plots, histogram charts, bar charts (besides histograms), line charts, and pie charts. Visual variables associated with representation of sentiment/emotion data were thus considered in relation to a particular encoding and data. Finally, we should mention that *interaction* was beyond our particular research questions for this study, and thus the respective tasks were designed for static visualizations only.

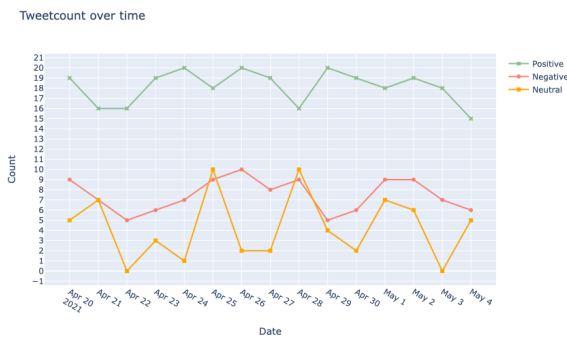


Figure 1: One of the study tasks with the question text *How many neutral tweets were posted between Apr 26 and Apr 28?* and answer options 18, 14, 10, and Don't know.

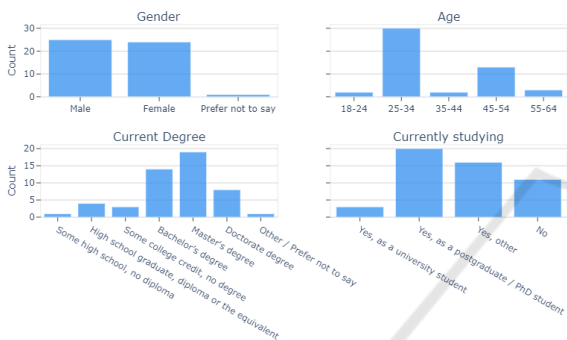


Figure 2: Responses for the demographic questions.

### 3.2 Data and Implementation

Our study used two datasets introduced in the prior works as the basis for generating the visualizations for particular representations and tasks. The first dataset is a subset of Amazon reviews (Nibras, 2019), which includes prices, reviews, and scores for cell phone items. The second dataset is a Twitter dataset (Mohammad et al., 2018a; Mohammad et al., 2018b), which consists of three different subsets. The first subset consists of tweets with a sentiment/intensity score between 0 and 1. The second subset consists of tweets categorized by one of seven different intensity classes, from *very negative emotional state can be inferred* to *very positive emotional state can be inferred*. The last subset classifies each tweet into eleven fine-grained subjectivity categories (emotions and further aspects of subjectivity): *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust*. Based on these datasets and the considerations discussed above, 26 unique plots were generated in total (see an example in Figure 1), and these were used for 28 questions (several plots were re-used for different tasks, e.g., *extremum vs retrieve*).

To facilitate the accessibility of this study targeted at the general public, a decision was made to implement it as an online tool, which could be used from

most web browsers on most platforms. The particular visualizations were implemented in Python and exported as static figures. For most visualizations, *Plotly for Python, Bokeh, and Matplotlib* were used, and for the data handling, *pandas* was used. The online survey itself was implemented with an open-source JavaScript library *survey.js* and deployed as a *DigitalOcean* web application using *React.js*.

In order to test the feasibility of both the survey questions/tasks and the technical implementation, a pilot study was run with three invited participants with no visualization knowledge. Based on their feedback, the implementation was adjusted and technical issues were resolved. Afterwards, the online survey and the respective instruction pages were made available online for the general public and invitations were sent out through several university mailing lists. The results of the study are presented next.

## 4 STUDY RESULTS

In this section, we focus on the user participation, particular task results, and the reported feedback.

### 4.1 Overview

The online questionnaire was run between May 12 and June 6, 2021, and the results include the responses from 50 participants. They were informed on the start page that the participation was fully voluntary and could be withdrawn at any time, while their answers and response times would be recorded for future use. Participants were able to contact one of the authors for any further information regarding the study. No personal information except for the demographic question responses was recorded, and no compensation was offered to the participants.

The initial demographic question responses are summarized in Figure 2. Next, most participants reported that they had limited (18 participants) or no (14 participants) experience with visualization; 47 participants reported no color blindness issues, while the other 3 participants reported mild issues; 35 participants took the survey on their computer or tablet, and the other 15 participants took it on their phones.

The participants were then asked to solve three warm-up questions and were warned that the answers and response times would be recorded afterwards. The responses for the next 28 single-choice questions, which were designed according to the considerations discussed in Section 3 and displayed to the participants in a shuffled order to minimize learning effects, are discussed in the following subsection.

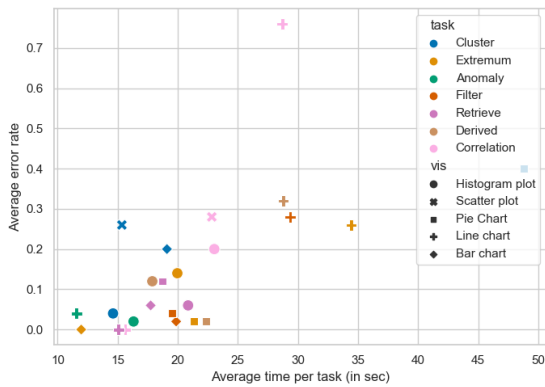


Figure 3: Average time per task vs. the average error rate per task. The task name is encoded using color and visual representation as the marker shape.

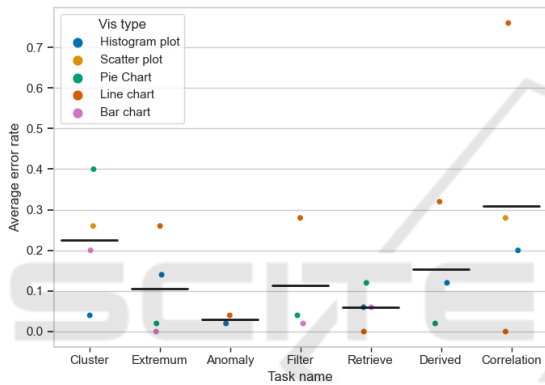


Figure 4: Different task types vs. the average error rate per task. Color encodes the representation. The horizontal lines represent the mean of the average error rates per task.

### 4.2 User Task Results

The majority of participants answered  $\approx 90\%$  of the questions right. The average total time spent by participant is  $\approx 10$  minutes, and the average response time is  $\approx 20$  seconds per question. Figure 3 visualizes the distribution of the average error rate and time per task. The question with the lowest error rate (0%) is *What phone brand has the lowest average rating?* (the *extremum* task / a *bar chart*), and it also has second-lowest response time ( $\approx 12$  seconds).

Figure 4 focuses on different task types vs. the average error rate. There is only one outlier here: a question from the *correlation* task (*From the above graph, can you tell that there is a (weak) correlation between ratings and the number of reviews?*) with a *line chart*. This may be due to the fact that this question was rather hard to be answered by the general audience. Participants were the most effective in the following tasks, sorted from most to least effective:

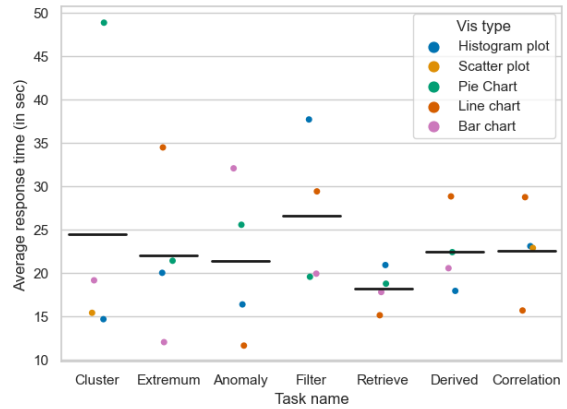


Figure 5: Average response time per task with representation encoded with color. The horizontal lines represent the mean of the average response times per task.

*retrieve*, *extremum*, *filter*, *derived*, *cluster*, and *correlation*. The *anomaly detection* task is left out here because it was represented by only two questions, which might not be sufficient to estimate its effectiveness.

Figure 5 demonstrates the average response time per task. The *clustering* task has a question with the highest average response time of  $\approx 48$  seconds. The fastest response time is observed from the *anomaly* task using a *line chart* ( $\approx 12$  seconds). Without the outlier of the *cluster* task, pie charts also tend to be around the average response time of 22 seconds. Similar can be observed for histogram plots. Regarding the task results in relation to the demographic information discussed above, our data suggests that the gender of participants neither had an influence on response time nor the correctness rate. The results for the educational status and device type are mixed, too.

### 4.3 User Preferences and Feedback

The final part of the study included several questions about the participants' confidence and preferences. 38 participants reported moderate or high confidence regarding their responses. Next, the participants were asked to rank the visual representations from easiest to hardest to work with, in their opinion<sup>2</sup>. The results for this question were somewhat heterogeneous, with 4 responses supporting the ranking of (*bar charts*, *histograms*, *line charts*, *pie charts*), 3 responses for three different rankings, and 2 or 1 responses for other rankings. The next question was about the color coding used in the user tasks. 36 participants stated that some colors were easier to work with than others, 10 stated that they did not perceive a difference in efficiency,

<sup>2</sup>Scatter plots were not included here as they were only used in two tasks in the study.

and 3 stated they did not know, while one participant skipped this question. The next multi-choice question was *What color (in general) would you find best for identifying positive and negative emotions?* with several proposed combinations as well as the *Don't know* and *Other*. Here, 46 participants stated that their color preference is *Green for positive and red for negative sentiment*; 4 participants chose *white for positive and black for negative*; 4 participants stated other preferences, and two of them wrote afterwards: *"I don't care as long as it's consistent over all plots"* and *"Green for Positive, Black for Negative. Not as intuitive, but looking better on a screen."*

As part of the final free-text remarks, we recorded the following responses, among others:

- *"Didn't understand the usage of question 41 [preferred ranking of representations, see above] initially and didn't find a way to correct it. Correct order would be 1: Bar Chart, 2: Pie Chart, 3: Line Chart, 4: Histogram Chart. Also, a small visualization at this point would help tremendously to recap which was which. In the question about what was looking odd in the pie chart, I missed the option 'only one text was white, the others black'."*
- *"Having to look at and deal with pie charts is a cruel and unusual punishment."*
- *"In general it was a good survey. When I was using my phone for the first attempt I misclicked and wanted to go back, but there is no 'previous' button. It's very quick shift between questions, I wasn't very sure that it registered the answer I chose."*

## 5 DISCUSSION

In this section, we discuss the outcomes and limitations of our study, which might provide guidance on the design of sentiment visualization approaches and inspiration for further research on this topic.

### 5.1 Study Outcomes

By analyzing the results from a different perspective (see Figure 6), we can note that the users were most effective with the following representations (in decreasing order): bar chart, histogram plot, pie chart, line chart, and scatter plot. However, we should note that the survey questions only included two scatter plots. Thus, the observation of the effectiveness of that particular representation might not be reliable.

By investigating Figure 5, it is clear which tasks seem to be answered with shorter response times,

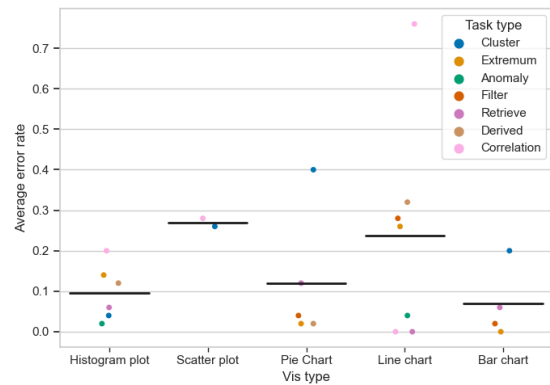


Figure 6: Average error rate per question categorized by representation. The horizontal lines represent the mean of the average error rates per representation.

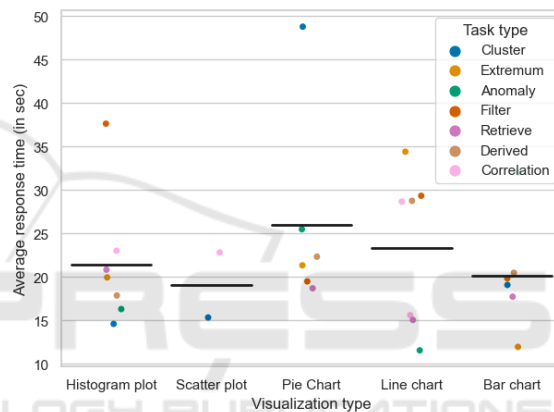


Figure 7: Average response time per question categorized by representation. The horizontal lines represent the mean of the average error rates per representation.

thus being more efficient than others in the context of sentiment visualization. The simpler tasks regarding the average response time in ascending order are *retrieve*, *anomaly* (note: used in two questions only), *extremum*, *derived*, *correlation*, *cluster*, and *filter*. With a similar approach, we get the most efficient representations from Figure 7 (in descending order): scatter plot (note: used in two questions only), bar chart, histogram plot, line chart, and pie chart. To get to a more fine-grained analysis of efficiency, we take a look at some specific questions. The question which took the participants the longest to answer is from the *cluster* task, *What are the three main clusters of emotional states?* using a pie chart, with an average response time of 48.6 seconds.

For the efficiency of visual variables, we observe similar patterns as with effectiveness. Users tend to identify the polarity of sentiments faster when we used green for positive and red for negative emotion. In terms of neutral polarity, we could not iden-

tify differences in efficiency. However, many participants noticed graphs where polarity seemed to have a “wrong” color. This means that they had expected that some sentiments should have another color than the one used. As we used many different colors in our graphs, there is no clear correlation between a particular color encoding and other less/more efficient visual channels. We thus come to the following conclusion with regard to channels, ordered from most to least efficient: color, bar size, and point position.

As our survey also showed some interesting results on user satisfaction (Frøkjær et al., 2000), preferences, and free-form user feedback, we also take some short notes on these results. First of all, most of the participants stated that they were moderately or very confident in answering the questionnaire. The free-form feedback for the technical implementation and contents of the survey was overall very positive, and only a few statements were made about some confusing design choices. Regarding the preferences for particular colors, many users stated that they found it easier to work with green/red and light blue colors for positive, negative, and neutral emotions. Only a minority stated that color did not influence their performance and satisfaction. Thus, we suggest considering this color scheme for sentiments/emotions (while keeping color blindness concerns in mind).

It is also interesting to notice that the participants identified the polarity of the sentiments/emotions of *surprise* and *trust* differently. Only 11 participants regarded these two sentiments as positive, while 36 participants stated that only one of these sentiments have positive polarity and the other neutral polarity. 28 participants stated that *trust* has positive, and 12 stated that this sentiment has neutral polarity. It clearly shows that every participant defines the polarity of these categories differently.

## 5.2 Limitations and Open Challenges

There are some limitations of our work which can be taken into account for further research efforts. First of all, the implementation of the study as an online survey rather than an experiment in a controlled lab setting might have implications for the reliability and reproducibility of the results (Fekete and Freire, 2020). We aimed to mitigate some potential issues by introducing several mandatory warm-up questions before the main time-tracked part of the study. Several minor technical issues were reported, for example, several participants had issues with the ranking question interface, which might have affected the reported results on the users’ preferences to some extent.

The participants in this study were asked to com-

plete tasks using static visualizations only. We encourage further research on the effectiveness and efficiency of sentiment visualization techniques while taking interactivity (Munzner, 2015) into account. The number of visual marks shown in the visualizations in this study was also limited to a rather small number, as it could require increased duration and complexity of the study. Future research should look at how the data point cardinality affects the users’ performance in the context of sentiment visualization. This could involve replication of this study with larger datasets, larger numbers of unique plots and questions, but also a larger number of participants. Empirical data concerning further representations, potentially involving additional visual variables/channels and even 3D representations, would also be valuable.

Finally, we should remember the concern discussed by Isenberg et al. about the evaluation of complex visualization / visual analysis approaches, which cannot be reduced to controlled experiments focusing on individual visual representations and low-level interactions (Isenberg et al., 2013). This challenge is still open with regard to sentiment visualization problems—thus, further evaluation of such complex and feature-rich approaches and solutions that involve sentiment and emotion data visualization (with the results and guidelines highlighted in this work in mind) remains as another opportunity for future research.

## 6 CONCLUSIONS

This study investigated effectiveness and efficiency of common visual representations and variables for various user tasks in the specific context of sentiment visualization. We conducted an online task-based within-subject user study with 50 participants from the general audience. The study involved static sentiment visualizations with several common visual representations and seven user tasks, resulting in 28 questions and further input on users’ preferences and feedback. After the analysis of the results, we discussed how different visual variables and representations influence the mentioned usability measures, and we also outlined design recommendations and opportunities for further research on this topic.

## ACKNOWLEDGEMENTS

This research has been partially supported by the funding of the Center for Data Intensive Sciences and Applications (DISA) at Linnaeus University. The authors would also like to thank the study participants.

## REFERENCES

- Alharbi, M. and Laramee, R. S. (2019). SoS TextVis: An extended survey of surveys on text visualization. *Computers*, 8(1).
- Amar, R. A. and Stasko, J. T. (2005). Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442.
- Bertin, J. (2011). *Semiology of Graphics: Diagrams, Networks, Maps*. ESRI Press. Translated by W. J. Berg.
- Bouchama, S. (2021). Task-based evaluation of sentiment visualization techniques. Master’s thesis, Linnaeus University.
- Boumaiza, A. (2015). A survey on sentiment analysis and visualization. *Journal of Emerging Technologies in Web Intelligence*, 7(1):35–43.
- Carpendale, S. (2003). Considering visual variables as a basis for information visualisation. Technical Report 2001-693-16, University of Calgary.
- Carpendale, S. (2008). Evaluating information visualizations. In *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950 of LNCS, pages 19–45. Springer.
- Chen, S., Lin, L., and Yuan, X. (2017). Social media visual analytics. *Computer Graphics Forum*, 36(3):563–587.
- Diakopoulos, N., Naaman, M., and Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, VAST ’10, pages 115–122. IEEE.
- Ebert, A., Gershon, N. D., and van der Veer, G. C. (2012). Human-computer interaction. *KI — Künstliche Intelligenz*, 26(2):121–126.
- Elmqvist, N. and Yi, J. S. (2015). Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269.
- Fekete, J.-D. and Freire, J. (2020). Exploring reproducibility in visualization. *IEEE Computer Graphics and Applications*, 40(5):108–119.
- Fekete, J.-D., van Wijk, J. J., Stasko, J. T., and North, C. (2008). The value of information visualization. In *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950 of LNCS, pages 1–18. Springer.
- Frøkjær, E., Hertzum, M., and Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’00, pages 345–352. ACM.
- Görg, C., Pohl, M., Qeli, E., and Xu, K. (2007). Visual representations. In *Human-Centered Visualization Environments: GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany, March 5–8, 2006, Revised Lectures*, volume 4417 of LNCS, pages 163–230. Springer.
- Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., and Möller, T. (2013). A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827.
- Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific Visualization Symposium*, PacificVis ’15, pages 117–121. IEEE.
- Kucher, K., Paradis, C., and Kerren, A. (2018). The state of the art in sentiment visualization. *Computer Graphics Forum*, 37(1):71–96.
- Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536.
- Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Meiselman, H. L., editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.
- Mohammad, S. M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018a). SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval-2018.
- Mohammad, S. M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018b). SemEval-2018 Task 1: Affect in tweets (AIT-2018). <https://competitions.codalab.org/competitions/17751>. Last accessed: November 8, 2021.
- Munero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.
- Munzner, T. (2015). *Visualization Analysis & Design*. CRC Press/Taylor & Francis Group.
- Nibras, G. (2019). (Un)locked cell phone ratings and reviews on Amazon. <https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews/version/1>. Last accessed: November 8, 2021.
- Purchase, H. C. (2012). *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press.
- Saket, B., Endert, A., and Demiralp, Ç. (2019). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512.
- Shamim, A., Balakrishnan, V., and Tahir, M. (2015). Evaluation of opinion visualization techniques. *Information Visualization*, 14(4):339–358.
- Zhao, J., Gou, L., Wang, F., and Zhou, M. (2014). PEARL: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, VAST ’14, pages 203–212. IEEE.