Linnéuniversitetet
Kalmar Växjö

Master Thesis Project

Task-Based Evaluation of Sentiment Visualization Techniques

Author: Samir Bouchama
Supervisors: Dr. Kostiantyn Kucher,
Prof. Dr. Andreas Kerren
External Supervisor:
Prof. Dr. Achim Ebert
Examiner: Dr. Narges Khakpour
Semester: VT 2021
Course Code: 5DV50E
Subject: Computer Science

Abstract

Sentiment visualization techniques are information visualization approaches that focus on representing the results of sentiment analysis and opinion mining methods. Sentiment visualization techniques have been becoming more and more popular in the past few years, as demonstrated by recent surveys. Many techniques exist, and a lot of researchers and practitioners design their own. 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. Multiple surveys and evaluations exist that argue for the importance of investigating the usability of such techniques further. This work focuses on evaluating the effectiveness, and efficiency of common visual representations for low-level visualization tasks in the context of sentiment visualization. It shows what previous work has already been done by other researchers and discusses the current state of the art. It further describes a task-based 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. This study is used for evaluating sentiment visualization techniques on their usability with regard to several sentiment and emotion datasets. This study shows that each visual representation and visual variable has its own weaknesses and strengths with respect to different tasks, which can be used as guidelines for future work in this area.

Index terms— sentiment visualization, sentiment analysis, sentiment evaluation, bar charts, visual variables, polarity, valence, user study

Preface

Thank you to all my friends and acquaintances for helping me to conduct the user study. Without you, I would not have that many participants taking part in it. Thank you to my parents, who motivated me throughout the time of writing this. You never ceased to motivate me.

Also, a big thank you to Dr. Kucher for being my primary supervisor for this project. Thank you for always giving me feedback and providing me with positive and supportive input on my work.

Contents

1 Introduction

This section introduces the topic of the thesis project and the current literature that focuses on evaluations and studies that analyze sentiment visualization techniques. In particular, it lists the limitations of the prior works and shows how this thesis project can contribute to the research in this area.

1.1 Background

General data visualization is widely used as part of various data analysis workflows [1], and it is supported to at least some extent by multiple data analysis tools. For example, several tools allow using static data visualization techniques for easier analysis of the underlying data. If graphs or diagrams are used, there is always a decision that has to be made on which visualization approach to choose. Numerous visualization methods and techniques are often described in related literature within the fields of information visualization [2] and visual analytics [3, 4].

More specifically, analyses of the use of such approaches for practical applications inside and outside of the academic environment [5–8] have been discussed in the literature. Additionally, guidelines for the choice of particular existing techniques and tools [9–11], and investigations and guidelines for the design of visualizations for a given task and data with the limitations of human perception and cognition in mind [12–16] have also been discussed in related works.

However, the focus of the research in this thesis project is on the part of visualizations that are used to visualize sentiment and emotion data.

We analyze various parts of these visualizations in greater detail. Different visualization types, also named *visual representations* or *visual metaphors*, are used to represent data in an informative way. Such representations can be, for example, bar charts or scatter plots. These charts make use of so called *visual variables* or *visual channels*, which are attributes that are used to encode particular quantitative or qualitative information values with low-level graphic elements of the charts (the elements themselves are known as *marks*) [17]. For example, size and color of a point in a scatter plot are visual variables/channels. We will go into more detail on this in Section 2.4.

As discussed by Munezero et al. [18], while the concepts of sentiment, emotion, opinion, and affect are not the same, they are related and are often used interchangeably as part of practical applications dealing with detection and analysis of subjectivity.

Such applications are manyfold nowadays, for instance, focusing on sentiments and emotions in social media such as blog posts, news site comments, shopping site reviews, etc. [19]. Text, video, and audio data can all be used to transmit information. This research, however, is limited to text data.

Detection and analysis of sentiments, emotions, or opinions in text data is part of the study within the fields of computational linguistics, natural language processing, and text mining [20, 21].

The particular problem (and the respective subfield) is known as sentiment analysis and opinion mining. Various approaches ranging from keyword matching to deep learning have been discussed in the literature to address the respective tasks [22–25].

The typical task is to classify the polarity/valence of a given text snippet (sentence, paragraph, document, etc.) as *negative*, *neutral*, or *positive*. However, there are

further alternatives with regard to the set of categories (e.g., several basic emotions, such as *angry* or *depressed* could be detected instead of polarity), scope and target of the classification, and so on [26].

In sentiment and opinion mining, sentiment visualization techniques are prevalent, with applications ranging from static presentation of classification results to interactive exploratory analyses. The majority of the literature focuses on static dataset visualization techniques. There are numerous techniques for visualizing sentiment and emotion data, and the possibilities for visualizing opinions are virtually limitless. However, only a few usability studies of sentiment visualization techniques in the literature address their efficiency and effectiveness or provide corresponding guidelines for their design and application.

1.2 Motivation

As mentioned above, the literature is full of numerous techniques for visualizing sentiment. At the same time, numerous surveys and evaluations of information visualization approaches exist, but they mostly focus on general-purpose data visualization approaches, with no specific evaluation for sentiment visualization.

The survey by Kucher et al. [27] classifies numerous sentiment visualization approaches according to the underlying data, designated technique tasks, and actual representation. However, that survey does not specifically focus on the evaluation approaches applied for the respective techniques. It can be noted that the amount of research demonstrating the efficiency and effectiveness of various sentiment visualization approaches is low. Here, *effectiveness* is defined by the accuracy and completeness with which users achieve certain goals [28]. *Efficiency* is the relation between the accuracy and completeness with which users achieve certain goals with respect to the resources taken, e.g., the time taken to complete the task [28]. Both of these measurements, alongside with users' satisfaction and other factors, contribute to the estimation of *usability* of human-computer interaction techniques [28], including information visualization. This is the primary motivation for assessing the effectiveness and efficiency of sentiment visualization techniques.

Shamim et al. [29] perform a usability analysis of the visualizations in several opinion mining systems. The primary objective is to rate different sentiment mining visualizations to examine perception discrepancies between different participants and assess critical visualization metrics [29]. The study of Shamim et al. thus relies mainly on the preferences and satisfaction reported by the users, with factors such as "eye pleasing" or "easy to understand" involved. Understanding the users' preference of a particular visualization technique, on the other hand, is insufficient for selecting a successful technique.

Another motivation for this work is the lack of literature on task-based evaluations [30] of such visualization techniques. These evaluations assess the usability of various visualization approaches for different analysis tasks. For instance, one chart type may be better suited to answering a specific type of question than others.

There are several important aspects of sentiment visualization techniques to consider with regard to their purpose and design. Firstly, they are used to represent the polarity or emotions of a piece of (textual) information. Polarity of information is, for example, *positive*, *neutral*, or *negative*. Examples of emotions are *angry*, *happy*, *sad*, *excited*, etc.

Another important aspect is the concept of aggregated sentiment and opinion data. Usually, longer text or sequential information is composed out of many different emotions such that it is required to aggregate sentiments. An example of this is to visualize the accumulated or average sentiment over time for a collection of social media posts.

Both of the previously mentioned aspects of sentiment visualization are fundamental in analyzing such techniques. Thus, they are important and play a significant role in our work.

It also assists in effectively selecting visualizations for various analytic and visual tasks. Due to the lack of evaluations on sentiment visualization techniques, practitioners might select non-optimal representations from general visualization research. For example, a pie chart is a well-known representation available in a large number of software tools. Still, it is known for being prone to issues when not approached carefully [31] even for general-purpose data. It might be even more problematic for representing and analyzing sentiment data in particular.

1.3 Problem Statement

The existing literature lacks sufficient empirical evidence about the effectiveness and efficiency of visual representations for sentiment visualization techniques. There is an insufficient amount of reference available for practitioners that could be used for finding the best technique for their specific needs. This is a critical state, as numerous sentiment visualization techniques have emerged over the past decade, and many more will emerge as a result of the growing interest in this area. This work would fill a gap in the literature by providing an overview of the usefulness and usability of various visual representations concerning particular user tasks in the context of sentiment visualization techniques.

1.4 Research Questions

This thesis project aims to address the following research questions in relation to the research problem described above. Table 1.1 lists both of the research questions of this work.

Table 1.1: Thesis project research questions.

RQ1	Which visual variables/channels are most effective and efficient with regard to visual encoding of polarity and emotions?
RQ2	Which visual metaphors/representations are most effective and efficient with regard to visual encoding of aggregated sentiment or opinion data?

To address **RQ1**, comparative evaluation [32, 33] of several encodings using alternative visual *variables* [12] will be necessary, for instance, indicating polarity with color (green for positive and red for negative polarity) vs. filled area amount for a bubble chart [34].

To address **RQ2**, comparative evaluation of several encodings using alternative visual *metaphors* [17] will be necessary, for instance, indicating the accumulated or average sentiment over time for a collection of social media posts with a line plot vs. a stacked area chart [35].

1.5 This Thesis Report

This report focuses on the following parts on the evaluation of sentiment visualization techniques. We discuss these areas in later sections.

- Review of related work
- Experimental design of a task-based user study
- Conducting the user study
- Analysis of the user study results
- Designing guidelines for practitioners

We aimed to investigate which sentiment visualization techniques can be applied to sentiment and emotion data as data visualization methods. We primarily conducted a literature review and a task-based user study, as discussed in Section 4.

We re-created some findings by Kucher et al. [27], as well as introduced a Twitter and Amazon reviews dataset from Mohammad et al. [36, 37] and Nibras [38], respectively. We visualized these two datasets and made design decisions for our experimental design. To accomplish this, we generated various sentiment visualization representations and posed some questions about various user tasks. The final task-based user study was the outcome of this experimental design. The primary goal of this user study was to keep it as straightforward as possible while still capturing the majority of aspects of sentiment visualization, such as how visualization types are used to represent sentiments in an informative way. The task-based user study sheds light on how practitioners should approach sentiment visualization techniques in their own practice. The user study's findings can be used to develop broad recommendations that point specialists in the right direction when determining which visualization technique to use. Additionally, this evaluation aims to advance the application of sentiment visualization techniques to similar data.

1.6 Contributions

The main contributions of this work are the findings of the task-based user study, which provide evidence about the efficiency and effectiveness of various sentiment visualization techniques concerning multiple user tasks, visual variables/channels, and visual metaphors/representations, as described in Sections 1.1–1.2.

These findings can be applied for choosing sentiment visualization strategies for communicating the results of computational and/or manual sentiment analysis methods.

Further feedback provided by the study participants and the particular guidelines based on the user study results are also part of the contributions of this thesis project.

1.7 Target Groups

This research is directed toward experts in opinion mining and sentiment analysis. In particular, specialists who employ various sentiment visualization techniques to accomplish their objectives may benefit from this work. Additionally, practitioners in need of guidelines for sentiment visualization with various common visual representations may find this work useful.

Finally, researchers in information visualization and visual analytics whose interests include sentiment visualization may use the findings of this thesis project for their research purposes.

1.8 Report Structure

This subsection briefly summarizes the layout of this work and provides a synopsis of our study. It shows how the report is organized and what it is about.

Section 1: Introduction

The first section describes the project and addresses the state of sentiment visualization at the time of writing. Further, it states the research problem and different research questions.

Section 2: Background

Section 2 defines sentiment visualization and summarizes the material that is already available. It summarizes the various assessments and surveys that are currently present in the literature. It also covers several static sentiment visualization approaches introduced by different authors.

Section 3: Methodology

This section introduces the actual task-based user study that was conducted in this work. It lists the different tasks and visual representations that were implemented. These are the main design choices that were used to conduct the user study. It also demonstrates some of the approaches used to interpret and assess the data.

Finally, this section includes a discussion of the reliability and validity of this study, as well as the ethical considerations involved.

Section 4: Experimental Design

This is the main section of this thesis, where we present how we implemented the user study. It addresses the experimental design decisions we took during implementation. It details each task and question and briefly explains how they were chosen. It examines how these questions may affect the study's results, final analysis, and evaluation. Further, it discusses some technical aspects of the implementation and how participants were chosen for the study.

Section 5: User Study Results

This section thoroughly analyzes and reviews the user study's results. It explores how the user study's findings relate to the subject and how they can be expanded onto. We introduce several tables and figures that show how participants performed.

Section 6: Discussion

This section details the methodology used to determine the effectiveness and efficiency of sentiment visualization representations. Additionally, it provides some broad and precise guidelines designed to assist other researchers in choosing visual-

izations that suit their needs. Finally, it states how the study issues are addressed and how it answers the research questions.

Section 7: Conclusions and Future Work

The final section of the thesis summarizes and reviews the major findings of the results report. Additionally, it discusses the work's shortcomings and suggests possible work that could be done in the future.

2 Background

Visualization techniques have a wide range of applications for data analysis and dissemination purposes [1]. They are the subject of research in the field of information visualization; more specifically, it is concerned with interactive visual representations of abstract data and their role in the respective applications, as discussed by Card et al. [2]. Sentiment visualization is a particular area of research within information visualization.

Figure 2.1 shows eight examples of different sentiment representations. The figure is a screenshot of the `SentimentVis Browser` introduced by Kucher et al. [27]. We will not go into great detail about these particular approaches; instead, we mention them to provide a quick overview of the subject and demonstrate the breadth of the design space of visual encodings for sentiment visualization. Karduni et al. [39] develop a visual analytic system to aid the analysis of misinformation shared on social media. Kulahcioglu and de Melo [40] analyze the impact of color choices on word clouds and introduce guidelines on how to choose fonts and colors to facilitate the interpretation. Watson et al. [41] present an interactive visualization tool for script-based stories. Chamberlain et al. [42] show a novel visualization approach for visualizing sentiment and stance data. Cuenca et al. [43] introduce a multiresolution streamgraph approach to explore hierarchical time series. Fu et al. [44] develop a visual analytic tool to interactively explore user groups in a forum. Gomez-Zara et al. [45] show a system that recognizes the major characters in news articles, and that evaluates whether they are being depicted as heroes, villains, or victims. Harris [46] presents a type of network to detect and classify emotions in a set of news articles.

Figure 2.1: Eight different types of sentiment visualization techniques. The image is a screenshot of the `SentimentVis Browser` introduced by Kucher et al. [27]. The corresponding publications are from the following authors sorted from left to right and top to bottom: Karduni et al. [39], Kulahcioglu and de Melo [40], Watson et al. [41], Chamberlain et al. [42], Cuenca et al. [43], Fu et al. [44], Gomez-Zara et al. [45], and Harris [46].

During the past decade, sentiment analysis and visualization tasks have grown in popularity. This is because the amount of data that needs to be processed and

visualized is increasing quickly. These pieces of data consist of social media posts, consumer feedback, articles on news websites, or even whole blog posts. As the name implies, sentiment analysis studies the sentiment associated with a piece of knowledge. It aims to provide answers to the questions such as "How people feel" or "What is the users' opinion" about something [22].

Sentiment visualization is the task of visualizing the various sentiments discovered through sentiment analysis and opinion mining approaches, or, in some cases, representing the sentiments, opinions, or emotions specified explicitly, without the need to conduct computational analyses [27].

Most of the existing work in sentiment visualization is associated with subjectivity expressed in text data (i.e., authors expressing their subjective opinions rather than objective, factual statements) and thus typically identified (and quantified) with natural language processing (NLP) and data mining (DM) approaches. Such data mining techniques review text to infer user likes, dislikes, and sentiments [20, 21].

The earliest basic techniques for sentiment visualization come from the need to communicate the results of such computational approaches, even with basic well-known charts, text lists, or tables. On the other hand, current state-of-the-art methods make extensive use of novel visual analytic and data visualization methods to facilitate complex sentiment analysis tasks [27]—similar to the overall state of the art in text visualization [47–49], which subsumes the sentiment visualization subfield. The areas of computational and interactive analyses of textual data have grown in popularity over the past few years as more people share their views and thoughts online. Chen et al. [50] discuss how the growth of social media influences visual analytic techniques. When social media and microblogging rise in popularity, the data collected by these platforms grow as well and requires support from both computational and visual analytic methods to provide meaningful and useful insights to the interested users [50].

The remainder of this section discusses the main concepts and methods of sentiment analysis, as well as some well-known visualization techniques that are often used in the sentiment visualization space. See Figure 2.2 for a summary of the number of sentiment visualization techniques that appeared in the literature between the years 2001 and 2017. Additionally, this section describes further related work, focusing on sentiment visualization evaluations and their various outcomes.

2.1 Sentiment Analysis

Sentiment analysis is the method of estimating subjectivity—typically, negative/neutral/positive polarity—in text data [22, 26]. For example, companies regularly use it to track emotion in social media data, assess brand images, and better understand consumers. Consumers express their opinions and emotions more freely than ever before. Thus it is expected that text mining will become an indispensable method for monitoring and understanding their emotions and feelings. By automatically monitoring consumer reviews, such as poll results and media platform interactions, companies will understand what makes consumers excited or unhappy, allowing them to adapt goods and services towards their clients' needs. Since the emergence of machine-readable corpora, the research in linguistics has shown that language is not just a medium of communication knowledge about reality, but rather to decide what we are discussing, taking a position, and expressing our thoughts and emotions [51].

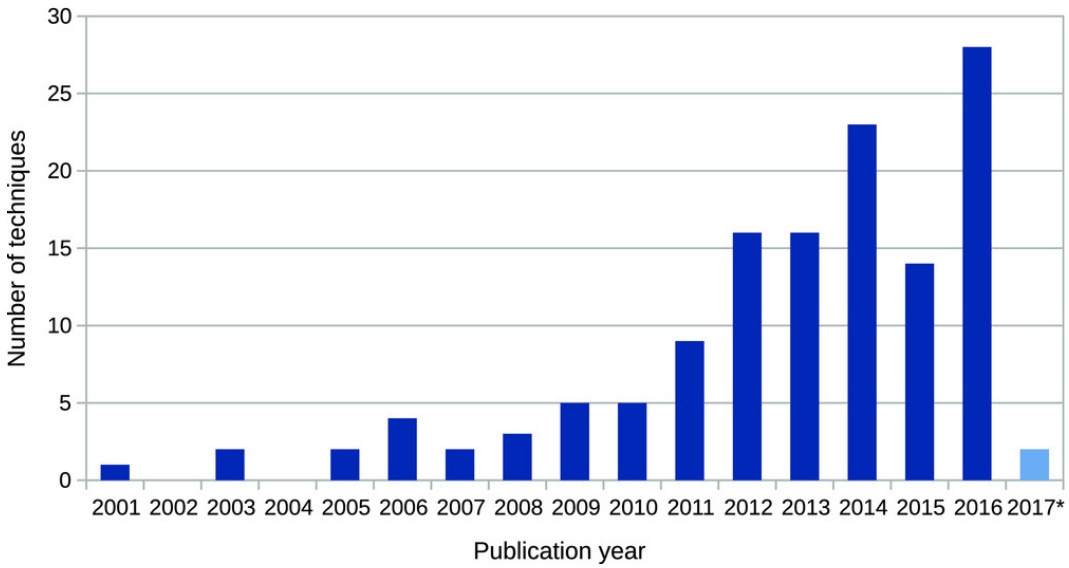


Figure 2.2: The growth of appearing sentiment visualization techniques between the years 2001 and 2017 [27].

Language usage in various ways is widely opinionated, collaborative, and intimate. Human contact serves a need. It is in continuous change, as is the use of expression.

Evaluative definitions are challenging to predict since they are not limited to standard vocabulary fields or basic terms but can be articulated across sections of words or larger chunks. In text mining or text analytics, opinion analysis (or opinion mining) is typically characterized as categorizing regions of textual data with regard to the topic/aspect and subjectivity expressed. This includes everything from essential words, phrases, and sentences to whole texts into a limited number of groups reflecting various emotions. However, where a more precise definition is needed, the term *sentiment* can be used interchangeably with words such as *mood*, *affect*, *emotion*, etc. as part of practical applications. As mentioned earlier, Munezero et al. [18] establishes the relation between these various concepts and argues that they are not the same, but acknowledges the often taken approach of their interchangeable use. We are involved in identifying highly emotional data or separating a positive sentiment from negatively loaded content. Subjectivity detection is closely related to one of these former tasks, while the latter is known as polarity detection [52, 53]. The grouping process is typically centered on the syntactic alignment of words from an originally described lexicon, for example, `SentiWordNet` by Baccianella et al. [54].

2.2 Types of Sentiment Analysis

As mentioned above, the most commonly used formulation of the sentiment classification problem involves three categories/classes of polarity, namely, *negative*, *neutral*, and *positive* [26]. Highly elaborate versions of the sentiment analysis task provide a wider variety of alternative sentiment models involving a continuous scale or an ordinal scale with more fine-grained categories, for instance, weakly negative to neutral to strongly positive [23]. A combination of sentiment analysis with other NLP and ML technologies, such as topic detection and tracking (TDT), is possible. For example, using sentiment analysis to automatically evaluate $4000+$ reviews online regarding the performance of a particular product brand could help a company

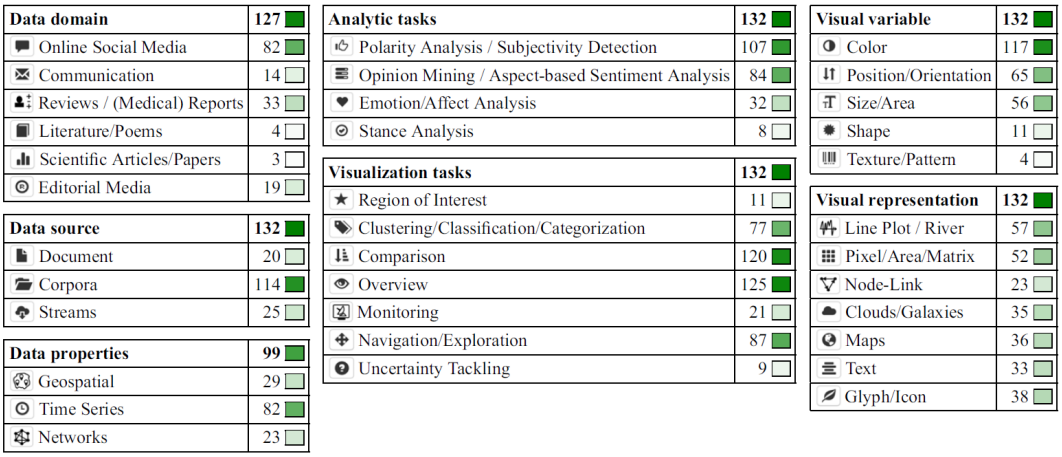


Figure 2.3: Different categories of sentiment visualization techniques discussed by Kucher et al. [27].

determine whether consumers are comfortable and satisfied with the pricing system and customer support. Similarly, a company might want to monitor the current brand comments on social media over time to identify unhappy customers easily. The applications of sentiment analysis are potentially endless and unlimited. Text analysis models concentrate not only on polarity, but also on emotions and attitudes (*angry*, *joyful*, *depressed*, etc.), urgency (*urgent*, *not urgent*), and more.

2.3 Sentiment Visualization Techniques

As mentioned, the literature is full of different sentiment visualization techniques. Due to its diversity, there is no single best strategy for all tasks. The following section summarizes some well-known and commonly used techniques from the literature that will be applicable in this work. One of the important sources is the work done by Shamim et al. [29]. In their work, they evaluate different visualization techniques gathered from a systematic literature review. The most used techniques can be categorized into the following representation groups: radial plots, graphs, hierarchical plots, and bar charts. This list is still manageable; however, the literature is full of many more different variations and combinations.

An interesting task in the sentiment visualization area is to find good representation techniques that represent the underlying data well and are easily understandable by the user. However, the work on this is limited to analyzing and evaluating such techniques. Thus, it is hard for practitioners to find the best method for their needs. A survey by Kucher et al. [27] classifies different techniques in different groups with fine-grained subcategories. The authors group these methods into multiple larger groups, consisting of the data domain, data source, and data properties. Two other groups deal with the task they solve, consisting of analytic tasks and visualization tasks. The last two groups describe what visual properties the visualization has. This includes what visual variables and representations are used. These larger groups consist of more fine-grained subcategories. For example, the visual variable group consists of the following subcategories: color, position/orientation, size/area, shape, and texture/pattern. The authors categorize 132 different sentiment visualization techniques into the aforementioned larger groups and subcategories. They further introduce an interactive survey browser that supports this categorization. It also allows authors of related research to publish their sentiment visualization approach

and categorize it similarly. This survey could help authors and practitioners to understand their and other related sentiment visualization techniques. An overview of these categories is shown in Figure 2.3.

Data aspects of the visualization techniques focus on the underlying data used, as discussed in particular by Kucher et al. [27]. Almost every visualization technique comprises an underlying high and low-level data aspect, which eventually affects the visualization pipeline's late stages. For instance, the original domain is an example of higher-level data, whereas the source representation is an example of a lower-level data aspect. A variety of sentiment visualization techniques are designed concerning particular data domain applications. Online social media, which includes forums, websites, weblogs, and social networking services, are a striking example of a media that contains a large amount of valuable textual information for sentiment analysis [50]. Surprisingly, a vast majority of systems in the data, amounting to over 62%, fall into this group in the survey by Kucher et al. [27]. Early examples of such techniques include the Mood Views NLP framework, which directly uses mood tags on blogs and websites as provided by clients' visitors on those sites and has so far suggested over a hundred mood views [55]. The approach uses simple line graphs for linear visualization. It encourages users to explore salient words and phrases for chosen times. The development of microblogs results in the emergence of more techniques focused on this type of data [56]. TwitInfo is an example of a sentiment visualization tool that was popular between the years of 2011 and 2014. As part of event analysis, Adams et al. [57] employ color-coded markers to denote the polarity of tweets. Another prominent example related to social media data is a tool titled Vox Civitas by Diakopoulos et al. [58] that facilitates journalistic inquiry of a US presidential address and the respective public reactions and discussions on Twitter. The overall sentiment detected in tweets for each minute is categorized as positive, negative, neutral, or controversial (if both positive and negative opinions were expressed prominently). Then represented visually using a colored bar timeline (similar to a stacked bar chart covering the complete timeline length). One more example relevant to social media data is a sentiment visualization approach introduced by Kucher et al. [59]. They introduce a visual analytics tool, titled StanceVis Prime, that is used for sentiment and stance [26, 51] analysis in temporal text data gathered from different social media platforms. The visual representations used for representing sentiment and stance in this tool include line graphs, bar charts, and glyphs.

However, social media is not the only possible source of data suitable for sentiment analysis and visualization—for example, Abbasi and Chen [60] discuss a visualization technique for text data from various data sources, including social media / online forums, email communication, and customer reviews. Their approach makes use of a bubble metaphor for the visual encoding of sentiment.

Text analytics and graphics are often used on email and chat discussions. This data element has been employed from over 10% of visualization techniques, according to the study by Kucher et al. [27]. For personal conversation history, an early apparent visualization approach was employed to depict individual messages as circles, organizing them in a three-dimensional layer based on technical ordering and discussion style [61]. Emotional signals are usually encoded using this method. By employing observed emoji symbols as the backdrop plane color, the approach captures the emotional content of talks. Another way for communication messages is an unusual emotion visualization approach that falls between computer graphics and

information visualization. The authors make use of a facial action encoding system (FACS) to build animated $3D$ avatars. Its faces represent the emoji icons associated with the corresponding text information. A new study on textual interpretation showed that chat logs containing emoticons might be categorized in polarity. Mail data is supported by inkblots and depicts the work of summarization by Guzman et al. [62].

2.4 Visual Variables and Representations

In this section, we briefly mention the concepts of visual representations/metaphors and variables/channels. Visual representations range from simple $2D$ visualizations to complex $3D$ graphs. In this case, the objective is to provide visualization tools that allow the user to comprehend and interpret the data by offering interaction methods for effective and efficient data communication [17]. Thus, visual representations are dependent on the context of their usage. As Görg et al. discuss [17], there are multiple different visual metaphors available—from simple and complex over to univariate and multivariate representations. Common visual metaphors are for example $2D$ plots, such as bar charts, scatter plots, pie charts, line charts, and many more. Certain representations are typically used for representing specific data types, for instance, line charts and stacked area charts [35] are often used for temporal data, while node-link diagrams and matrices are used for tree and graph/network data [34]. There are also theoretical design considerations and empirical evidence suggesting the feasibility (and usability) of particular representations for particular tasks: for instance, Görg et al. [17] provide an example of the same data represented with a pie chart and a bar chart, where the pie chart representation would not allow the users to discover differences in numerical values, in contrast to the bar chart.

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 [17]. Bertin and Berg [63] define the following seven visual variables: *position, form, orientation, color, texture, value*, and *size* [63]. These variables are distinguished perceptually without the use of cognitive steps in contrast to comparing written numbers, for example [17]. By storing data in a form that distinguishes between visual aspects, visual variables function as a communication medium. However, choosing the right visual variables hardly depends on the underlying data that is to be visualized. Depending on the data some visual variables are harder to work with than others [12, 34]. Munzner [34] 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*, in this context meaning that the importance of the data attribute should match the saliency/noticeability of the visual channel [34]. These considerations will play an important role in this work.

2.5 Evaluation of Information Visualization Approaches

Evaluation is typically mentioned in information visualization and visual analytics 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 [33, 64]. The choice of terminology is different in some cases in human-computer interaction, where “evaluation” (focusing on estimating the usability and

collecting mainly users' feedback for formative purposes as part of an iterative design-implementation-validation process) is contrasted to "experiment" (focusing on a single study event with the purpose of collecting empirical data in a controlled environment), as discussed by Purchase [32]. Evaluations of general purpose visualization and specific sentiment visualization techniques from recent years exist, as discussed, for instance, by Isenberg et al. [30]. This subsection discusses some works relevant to this topic in detail.

We can start this discussion by introducing challenges in information visualization evaluation. For this, we base ourselves on the points mentioned by Carpendale [65]. It is hard to conduct research when choosing the right focus and to ask the right questions. Also, in regard to this work, it is difficult to choose the right methodology to be sufficiently precise in procedure and data collecting. Most of this empirical research relates to human-computer interaction (HCI) research [32]. For example, many tasks in HCI are related to interface interaction, such as zooming, filtering, and accessing data details. Many other challenges are shared with HCI empirical research, but they would be out of the scope of this discussion. Another issue is that research software and this user study rarely reach the point where it can cover the entire range of tasks or be completely deployable in real-world conditions. Mainly in information visualization, many tasks vary with data type and character from low-level to high-level tasks. Some of these tasks, particularly those involving obtaining a new understanding of the data, are not well defined, making them more difficult to assess empirically. Examples of low-level detailed tasks also present in this work are compare, cluster, correlate, and categorize. High-level and more complex tasks that require an understanding of data trends are not present in this user study. Some examples of this would be learning causal relationships or predicting the future. Another point is that the results of such studies also strongly depend on the participants' motivation and their interest and knowledge about the questioned domain. There are many more challenges in evaluating information visualization, but this briefly illustrated that some research has already been done and will continue to be relevant in the future.

Elmqvist and Yi [33] propose a pattern-based approach for the evaluation of general data visualization. In total, they present 20 different patterns, which are also available in a Wiki which can be used by users and practitioners. The use of these so-called visualization evaluation patterns is to capture proven solutions to common problems in a form that can be reused that is accessible to non-experts. The authors thus, provide a catalog of best practices that further researchers can easily adopt in their own work. They specified their patterns in five basic components. Consisting of a name, a problem description, a solution on how to solve the problem in a reusable and flexible way, consequences when applying the pattern, and they each list examples to illustrate how to use it. They describe each of the 20 patterns in detail and classify them by the evaluation methods, whether they are quantitative or qualitative. The quantitative evaluation focuses on collecting performance measurements, and the qualitative evaluation collects more in-depth and free-form data. The patterns can be categorized into five main categories based on the high-level purpose that the researcher is trying to achieve. These five categories consist of *exploration*, *control*, *generalization*, *validation*, and *presentation* [33]. Overall the authors did a thorough study in gathering good patterns, which are also well known in the community. Concluding, the authors state that their work provides a powerful methodology for looking at evaluation and spreading experience as a whole. The limitation of this

work is that their article is not exhaustive and is limited to their own work [33]. The background of this could nonetheless also help in sentiment visualization evaluation.

Another strategy of visualization evaluation is introduced by Wall et al. [66]. The work describes a methodology for an evaluation approach to estimate and quantify the potential value of visualization. The overall work focuses on the assessment of visualization *value*, as formulated by Stasko [67]. The authors' procedure is to define several heuristics that should later be rated by domain experts on a Likert scale. They come up with 21 low-level heuristics. The methodology assessment was with visualization experts who were asked to use these heuristics to rate three different visualizations. From assessing the participants' ratings, one visualization clearly got the highest average score [66]. Thus, the evaluation clearly works, as stated by the authors. The analysis of this study's results shows that the evaluation methodology can be used for identifying the value of visualization. Still, it remains unclear if this methodology is applicable to other approaches. The study was only conducted on three different representations using the same dataset. Also, the evaluation methodology should be tested if it produces the right values. For this, it would be useful to compare it to so-called ground truth values [66].

Sentiment visualization has not gotten the same level of attention in extensive evaluations as other regularly associated visualization areas to data obtained from text, such as the study of topic templates or events. Very few text analytics so far have categorically employed sentimental or opinion analysis in this visualization. In recent work, Wanner and his fellow researchers did a survey on visual interpretation in text data streams [68]. The group selected a polarity extraction technique to perform text processing as a method of visual analytics. Fourteen of the fifty-one articles used in their study met the category. For example, in recent research, sentiment and effect analysis is a possible attribute extraction tool for text analysis. Despite the vast amount of research on sentimental analysis, there is still a gap to be filled in this field, and there is a need for more survey works. To the best of our understanding, only two published works comprise surveys on emotion visualization strategies. Recent research by Boumaiza [69] shows an overview of sentiment and viewpoint visualization strategies, emphasizing social media texts. However, the researcher does not have specific requirements or an appropriate classification. As a result, there is an overuse of various methods that are not specifically relevant to sentiment visualization, or even visualization in general, making it impossible to browse the survey and use it as a guide. According to our estimates, the work by Boumaiza [69] cites approximately 35 peer-reviewed sentiment visualization methodologies. An outline of 11 approaches and categories based on a visual metaphor. Still, the author did not classify the emphasis of the research [69]. Instead, it was done through an evaluation by performing an analysis to compare the approaches in terms of customer usability or usefulness. On the other hand, the research focuses on categorizing a much greater range of techniques in terms of various aspects relevant to the neural network, data, user tasks, and pictorial representation.

Another evaluation by Shamim et al. [29] reveals findings on the usability of sentiment mining systems' visualizations. They classify 11 techniques according to a visual metaphor. These techniques consist of, for example, opinion wheels, rose plot variations, bar charts, and more. Their work compares techniques concerning different metrics such as eye-pleasing, easy-to-understand, user-friendly, informative design, usefulness, and representation style. 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 others. They rank the top five study's visualizations as follows: bar chart, glowing bar, treemap, line graph, and pie chart. They also conclude that participants in their seminar, in which consumers were introduced to the topic, performed better than users who completed the online questionnaire.

2.6 Summary of the Related Work Analysis

This section summarized several evaluations and research on sentiment visualization techniques from the current literature. It demonstrates that numerous studies have been conducted on this topic. However, many methods have not been tested for their usefulness and productivity in assisting users with various tasks. This implies that further research on this subject is essential. Also, these studies could be a good starting point for our research. However, the current works often lack in visualization generality as often topic-based visualizations are used. To this end, the study will conduct additional evaluations on basic sentiment visualization methods.

3 Methodology

This section discusses the methodology followed for this thesis project. It briefly lists the steps we took during the work and how we planned each step. In later sections, we will go over the experimental design and implementation in greater depth.

3.1 Scientific Approach

The scientific approach followed in this thesis project is inspired by the empirical study design in information visualization [33] and human-computer interaction [32], with the main goal of establishing particular research questions of interest (see Section 1.4), mapping them to a particular experimental design, and collecting the respective empirical evidence that would allow for answering these research questions. The concrete methodology of this project consisted of (1) a literature research, followed by (2) designing, conducting, and analyzing the results of an online task-based user study of different sentiment visualization techniques. The survey was performed in a manner close to the work described by Saket et al. [70]. We designed seven different user tasks related to sentiment visualization. These tasks included, for example, finding erroneous data in a basic bar chart in the context of sentiment values data. To visualize sentiment and emotion, we used multiple sentiment analysis datasets depending on the task. Based on these visualizations, we came up with different survey questions for each plot.

3.2 Methodology Overview

The following list provides an overview of the methodology followed for this work. In the following subsections as well as Section 4, we explain the aspects of the experimental design and implementation in further detail.

- *Review of Related Literature*: in order to work on the experimental design, the analysis of the prior work had to be carried out.
- *Designing the Study*: the beginning of the experimental design consisted of coming up with an actual plan on how to design and conduct the study.
- *Datasets*: a critical part was to find relevant and easy to handle sentiment analysis datasets. Most of the datasets were retrieved from related work and will be further discussed in Section 3.2.3 and 5.
- *Defining Tasks*: this part consisted of defining several user tasks for the study.
- *Defining Visualizations*: for each task, a different sentiment visualization technique had to be defined. This included many simple chart types to ensure that they remained understandable to a broad audience.
- *Question Design*: the question design followed after formulating tasks and coming up with visualizations.
- *Online Study Implementation*: to conduct the study, the survey was implemented in an easy-to-use survey framework.
- *Pilot Study*: following the collection of visualizations, question, and task design, the questionnaire needed to be checked on feasibility and user-friendliness.

- *Finding Participants*: this task consisted of distributing and finding suitable participants for the online questionnaire.

- *Data Collection*: the data collection was performed via the online survey tool deployed for the study purposes as the participants provided their responses.

- *Analysis and Discussion*: the final part of the experimental design consisted of analyzing and discussing the results.

3.2.1 Review of Related Literature

The related work consisted of gathering already available evaluations, studies, and surveys. This part has already been discussed in detail in the previous sections of the background discussion in Section 2. We used the findings of the literature review as a starting point to design charts and questions. We noticed that several works used basic representation types as the main visualizations for their work. This led us to also use basic visualization approaches which are understandable by the general audience.

3.2.2 Task-Based User Study Design

The user study is targeted at the general public to facilitate the generalizability of the findings and to increase the potential number of participants. This requires that all tasks and visualizations should be intuitive to the average user, whose level of visualization literacy might be limited [71, 72]. The study is divided into three different parts: training, the main experiment, and follow-up questions, similar to the procedure typically employed for human-computer interaction studies [32]. During the training phase, participants are briefed about the purpose of the study and their rights. At this stage, the participants are asked some demographic questions, for example, age, gender, prior experience in creating/analyzing visualizations, etc. After this step, users are asked to perform different trial questions. These trial questions are used to familiarize the participants with the study setting. To prevent the participants from skipping the training questions, participants are not able to move to the next question unless they answer the previous question correctly. The main experiment section contains the actual study questions and tasks. After the training phase, participants are ready to start the main experiment. All questions of the main experiment part contain a "Don't know" option to skip the respective question. Participants are assigned questions in a random order to prevent users from extrapolating new judgments from previous ones. The follow-up questions are shown after completing the main experiment. The participants are asked to perform additional ranking and feedback questions. In these ranking questions, participants are asked to rank different visualizations and visual encodings in the order of their preference for performing a given task.

3.2.3 Datasets

The first dataset for creating visualizations is a subset of Amazon articles from Nibras [38]. It consists of ratings and reviews on Amazon (un)locked cell phone articles. Table 3.1 shows the first five data points of this dataset, it consists of the following features, *'asin', 'brand', 'title', 'url', 'image', 'rating', 'reviewUrl',*

'totalReviews', 'prices'. Due to space constraints, Table 3.1 shows only a relevant subset of features.

The second dataset is a Twitter dataset, consisting of multiple different social media posts (*tweets*). It was used in an online sentiment analysis competition and is from Mohammad et al. [36, 37]. It originally consists of three different subsets. Table 3.2 shows example data points of two of the three subsets of this dataset. 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. These classes define sentiments in different stages, 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). These include *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise*, and *trust*. Due to space limitations, we do not show this part in the table.

Both datasets are available online in a ready-to-use format. We only need one preprocessing step, as we only use a small subset of the data. The relevant preprocessing step consists of removing non-numeric values from the price column of the Amazon dataset.

Table 3.1: Example entries of the Amazon dataset by Nibras [38].

Brand	Title	$Rating$	$Totalreviews$	$Price(s)$
Nokia	Dual-Band / Tri-Mode Sprint PCS Phone w/ Voice Activated Dialing & Bright White Backlit Screen	3.0	14	NaN
Motorola	Motorola I265 phone	2.9	7	$49.95
Motorola	Motorola C168i AT&T Cingular Prepaid Gophone Cell Phone	2.6	22	NaN
Nokia	Nokia 6500 Slide Black/silver Unlocked Cell Phone	2.4	5	NaN
Motorola	Motorola i335 Cell Phone Boost Mobile	3.3	21	NaN

Table 3.2: Example entries of two subsets of the Twitter dataset by Mohammad et al. [36, 37].

ID	Tweet	Affect Dimension	Intensity Score
2017-En-30153	@user yeah! :) playing well	valence	0.6

ID	Tweet	Affect Dimension	Intensity Class
2017-En-30153	@user yeah! :) playing well	valence	0: neutral or mixed emotional state can be inferred

3.2.4 Task Design

Previously, Amar et al. [16] proposed a set of ten low-level analysis tasks that describe users' activities while using visualization tools to understand their data. First, these tasks are real-world tasks because users came up with them while exploring five different datasets with different visualization tools. Second, different studies used these tasks to evaluate the effectiveness of visualizations. With this in mind, we used a subset of the low-level taxonomy by Amar et al. [16], described below. This makes in total seven different tasks.

- *Find Anomalies*: Participants were asked to identify anomalies within a given set of data points with respect to a given relationship or expectation. These anomalies were manually created so that, once noticed, it would be straightforward to verify that the observed value was inconsistent with what would normally be present in the data (e.g., an absurd peak in a bar plot, which would obviously be considered as abnormal). An example question would be: *Does the data look abnormal*?

- *Find Clusters*: For a given set of data points, participants were asked to count the number of groups of similar data attribute values. For example, *How many clusters can you identify?*

- *Find Correlation*: For a given set of two data attributes, participants were asked to determine if there is a correlation between them. For example, *Is there a (weak/strong) correlation between sentiment intensity and time?*

- *Compute a Derived Value*: For a given set of data points, participants were asked to compute an aggregate value of those data points. For example, *What is the sum of all negative and neutral sentiments in the pie chart?*

- *Find Extremum*: For this task, participants were asked to find data points having an extreme value of a data attribute. For example, *What sentiment has the highest/lowest value?*

- *Filter*: For given concrete conditions on data attribute values, participants were asked to find data points satisfying those conditions. For example, *How many emotional states can you identify between time step X and Y?*

- *Retrieve Value*: For this task, participants were asked to identify values of attributes for given data points. For example, *Which emotion has green color?*

3.2.5 Visualization Design Choices

To generate visualizations, we used three pairwise combinations of three different data attribute types available in the datasets. In particular, nominal $\times$ numerical, ordinal $\times$ numerical, and numerical $\times$ numerical variable combinations. Figure 3.1 shows an example of each type. We did not include nominal $\times$ nominal because it is not possible to represent this combination using line charts. In general, to create scatterplots, bar charts, histogram plots, and line charts, we used the same length, font size, and color for drawing the respective axes, ticks, and labels. This does not hold for pie charts as they do not have axes. In addition, all the visual elements (for example, bars in a bar chars) used in the three charts had a different color. The main design decision that had to be made for pie charts was whether to include legends.

Instead of having legends, it would be possible to potentially add labels on the top of slices of pie charts. This, however, caused visual clutter, particularly in cases where the labels were long. Additionally, using legends for pie charts is a common practice in the majority of commercial visualization and table creation software, for example, Microsoft Excel, Google Docs, and so on [11]. The main decision was to not show any value on top of the slices of pie charts, instead showing the values of one data attribute using a legend and another one beside the slices. The colors used are the standard names from the CSS language used in HTML [73].

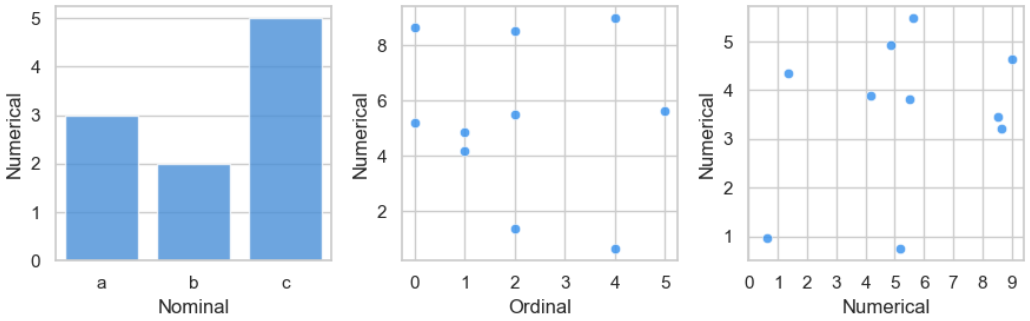


Figure 3.1: The three different attribute types used in our visualizations. From left to right, nominal × numerical, ordinal × numerical and numerical × numerical.

3.2.6 Online Study Implementation

The study was intended to be made available to multiple participants who might reside in different geographical locations and might prefer to take part at particular dates and times; furthermore, the participants could be using various operating systems or even using devices beyond a standard personal computer, such as tables or smartphones. Thus, to facilitate the accessibility of the study, a decision was made to implement it as an online tool, which could be used from most web browsers on most platforms. The complete implementation stack and the deployment process is discussed in more detail in Section 4.2.

3.2.7 Pilot Study

This process involved the preliminary testing of the feasibility of both the survey questions/tasks and the technical implementation. Questions were required to be easily understandable and answerable by the general public. For this, we selected 3 people with no visualization knowledge and asked them to take the study. In this step, no complaints were reported by the pilot participants, and the study was ready for the next step.

The next step consisted of evaluating the technical aspects of the study. For this we used the same three participants and got the following feedback. One participant had a few problems with the size of the visualizations. Consequently we updated the size of all visualizations. Another participant reported that the survey would start over when losing connection to the internet. This was also fixed, and after these changes, the study was ready for being published.

3.2.8 Finding Participants

This section talks about the distribution procedure of the study, which was carried out in several simple steps. At the beginning of the study, the survey link was distributed to friends, acquaintances, and university colleagues of the author of this thesis project. A short time later, the link was published in social media groups and distributed through an internal university mailing list at the University of Kaiserslautern to several students. The link was further distributed by the thesis project supervisor to his colleagues and acquaintances. The supervisor also helped with deployment of an informational HTML page in the ISOVIS group's website. This way, participants clicked on a link with a domain extension they were familiar and confident with.

3.2.9 Data Collection

During the study, all the different responses to the questions were measured. The study also collected response times for each question. Participants were informed of this technique at the beginning of the study. We tracked the completion times for better analyzing effectiveness and efficiency of different representation types and metaphors. This is a popular technique for gathering information about these usability measures [32, 33]. However, it is not common to find freely available web-based survey frameworks that already have this mechanism implemented, which was a concern during the prior implementation stage.

3.2.10 Analysis and Discussion

After analyzing the measures of task error rates and response times, we can use our findings to define multiple guidelines regarding sentiment visualization in the context of particular user tasks. The survey's follow-up questions, which ask about user preferences for different parts of the survey, are also crucial for this discussion. The information collected was not used in any way other than for the purposes indicated.

3.3 Reliability and Validity

The issues of reliability in information visualization research have been recently discussed by Fekete and Freire [74], with the recommendations to increase both (1) the computational reproducibility of the results with respect to the implementation and data, and (2) replicability of overall study results, especially in the context of user studies. The collection of data through a task-based study is a well-known method that is also presented in related work. However, it is obvious to say that when running the study multiple times, results will differ from each other, at least to some extent. For example, how fast and reliable is the internet connection of each of the participants? This may influence the results because the answering time for each question gets tracked. Some of the factors are, thus, unfortunately, outside of our control, and one possible solution would be to conduct a study in a controlled lab environment, which was hindered by the ongoing COVID-19 pandemic and the respective concerns. However, by discussing the experimental design and sharing the aggregated demographic data based on the participants' responses, we establish a foundation for replication of this study's results, which provides an opportunity for future work.

In the following, we discuss what validity definitions hold for this work. The following discussion briefly summarizes the different meanings of *validity* and how it links to this work. The definitions used originate from the website of the Department of Computer Science and Media Technology at Linnaeus University [75].

Construct validity is about the interpretation of theoretical constructs [75]. Section 5 provides an analysis of the results on *effectiveness* and *efficiency*, which are among the most important measurements of usability that our study focuses on. To make sure the terminology is interpreted correctly for the sake of construct validity, we address these terms again. As discussed by Frøkjær et al. [28], *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 [32]. *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 [28].

Internal validity is about establishing whether the results and conclusions follow the collected data [75]. According to Elmqvist and Yi [33], in the context of information visualization evaluations, internal validity can be achieved by controlling or eliminating parameters that are irrelevant to the research questions or tasks but might affect the obtained results regardless. In this thesis project, the data collection is treated as carefully as possible with respect to the chosen study procedure, i.e., an online task-based survey implemented with a custom interactive survey library. It should be mentioned, though, that some of the aspects of this study are beyond the experimenter's control, as discussed above. Additionally, some minor technical issues were registered during the ongoing study: for example, the ranking question on the follow-up page of the survey did not work for several participants. Thus, the results for the respective question cannot be claimed to conform to the definition of internal validity. This will be discussed in more detail in a later section of the report.

External validity is about establishing whether the generality of the results is justified [75]. As discussed by Elmqvist and Yi [33], in the context of information visualization evaluations, external validity can be achieved by the introduction of different study environments, participants, and real-world datasets. While the recommendation about several environments was not followed (as the study was only run once within the scope of this thesis project), the diversity among the participants was achieved by the fact that the survey was open to the general public for a relatively long time period. Furthermore, result analysis in Section 5 and Section 6 show that several domain experts took part in the study. Finally, the recommendation related to real-world data involvement was also followed, as both datasets used for the survey questions were based on real-world data and described in external sources, as discussed in Section 3.2.3.

3.4 Ethical Considerations

As this thesis project focuses on (1) collecting empirical data on sentiment visualization techniques usability and (2) analyzing the respective results and providing guidelines for visualization designers, we can discuss two aspects of ethical considerations relevant to this work.

First of all, we should consider the issues of ethics and privacy regarding the user study procedure and the respective data recording and storage. For this thesis project,

the procedure followed was inspired by the typical guidelines in the human-computer interaction field, as described by Purchase [32]. Participants were informed on the first page of the study that the participation was fully voluntary and could be withdrawn at any time, while their responses and response times would be recorded for future use. Participants were able to contact the corresponding author for any further information regarding the study. No personal information except for the demographic questions included within the survey was recorded.

With respect to the results of the study and the guidelines synthesized accordingly, we consider the responsibility of sharing the corresponding results as part of this thesis report. Further use of these results for various applications or future research by third parties is beyond our control, and it should be considered similar to the typical outcomes of publishing academic findings publicly.

4 Experimental Design

In this section, we elaborate on the design of the user study with regard to the particular questions/tasks, as well as the details of the online study implementation and some further concerns.

4.1 Question Design

This section explains question design choices and lists some example questions from each task of the study. As the question order is shuffled in the actual study, this list has no particular order. As color plays a big role in defining these visualizations, we used popular color names from the HTML markup language. For the full list of questions and visualizations generated, we refer to Appendix A.

The first example is a bar chart from the *retrieve* task. Figure 4.1 shows an example question of this type. This bar chart uses different visual variables. The colors, size, and orientation of the chart play a role. It shows the sentiment distribution of different tweets. Colors represent the emotional state of sentiment. Green represents positive sentiments, blue for negative, and cornsilk color for neutral emotions in this particular visual representation. The question to this graph is, *How many different emotions are shown in the graph?* with the options *9, 4, 11*, and *Don't know*. The correct answer is *9*.

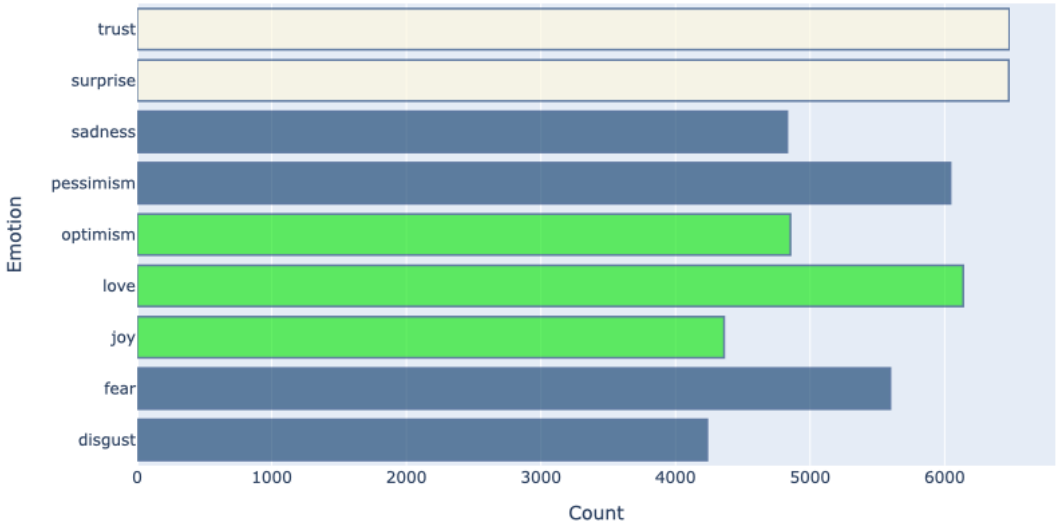


Figure 4.1: A bar chart showing different sentiments. It is color-coded in three different ways: green for positive, blue for negative, and cornsilk color for neutral emotions.

The next is an *anomaly* task, which Figure 4.2 shows with a histogram plot. This graph only consists of one color and defines bins for different intensity score values. The participants were asked the following question: *The above graph shows a distribution of sentiment intensity scores. At what point does the data look abnormal?* with options *0.8–0.85, 0.2–0.25, 0.5–0.55, The graph looks normal* and *Don't know*. The correct answer is *0.8–0.85*.

The next question consists of a *cluster* task. Figure 4.3 shows a scatter plot visualizing the count of tweets in relation to their intensity scores. This graph does not make use of various values of the visual variable of color for representing

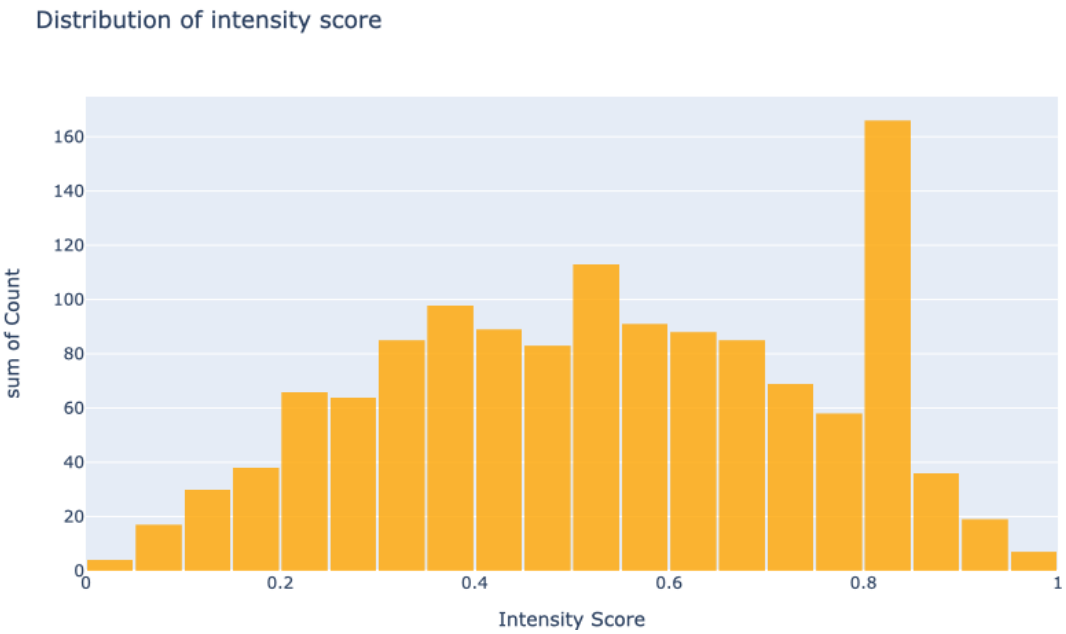


Figure 4.2: A histogram plot showing the distribution of the sum of count and intensity score of tweets. Color-coded in one way using orange color.

sentiment, as all of the dots are colored in the same way with green. The location of data points is another visual variable of this graph. The question asked in relation to this graph is: *Are you able to identify emotion clusters in the above graph?* with the possible options: *Yes, No*, and *Don't know*. The correct answer, in this case, is *No*.

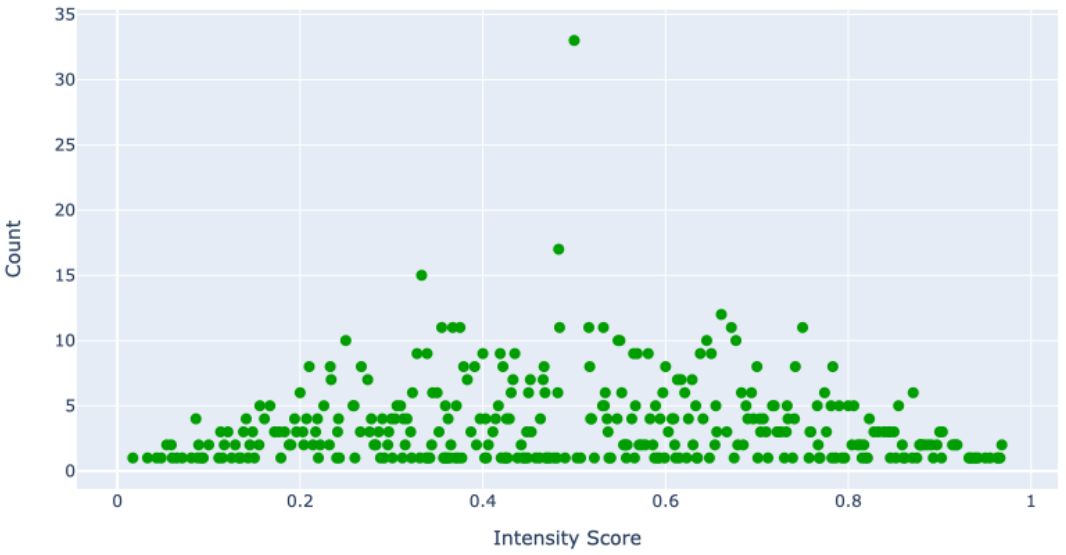


Figure 4.3: A scatterplot showing intensity scores of Twitter data. Location of data points is used for representing sentiment intensity scores.

The next example is a *correlation* task using a line plot shown in Figure 4.4. The graph shows the trend of emotion over time of Twitter tweets. This graph uses lines as marks, and the visual variable of color is used to indicate the respective emotion category. The different emotions use the following colors: orange for confident, royal blue for sad, dark sea green for energized, deep sky blue for pleased, tan for joyful, and purple for relaxed emotion. This graph was used in two different questions. The question concerning the *correlation* task was: *What type of emotion*

is constantly decreasing in count over time? with the following options: *Confident, Joyful, Energized, Pleased*, and *Don't know*. The correct answer is *Energized*.

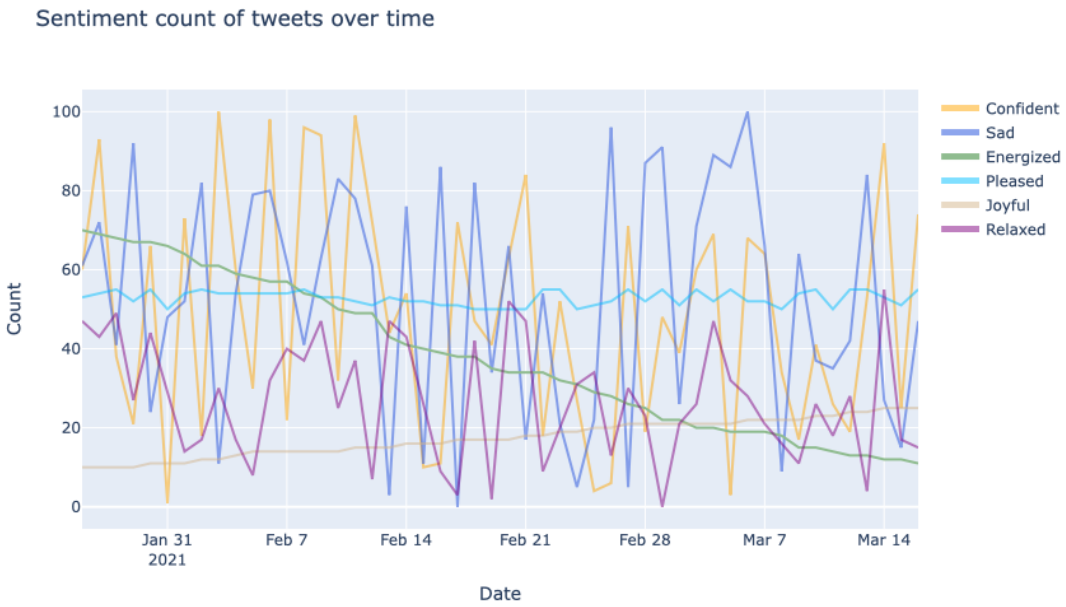


Figure 4.4: A line plot showing trends in emotions of Twitter data. Visual variables are the color and position of the points forming the lines.

The next task is to *calculate a derived value*. In this example, use a pie chart shown in Figure 4.5. The visual encodings of this chart are the area of the chart segments as well as their colors. The chart uses royal blue for negative, cyan for positive, and light cyan for neutral emotions. The question to this graph is *What is the total percentage of positive and negative sentiments without neutral sentiments?*, with the options: *61.4%, 71.1%, 28.9%*, and *Don't know*. The correct answer is *71.1%*.

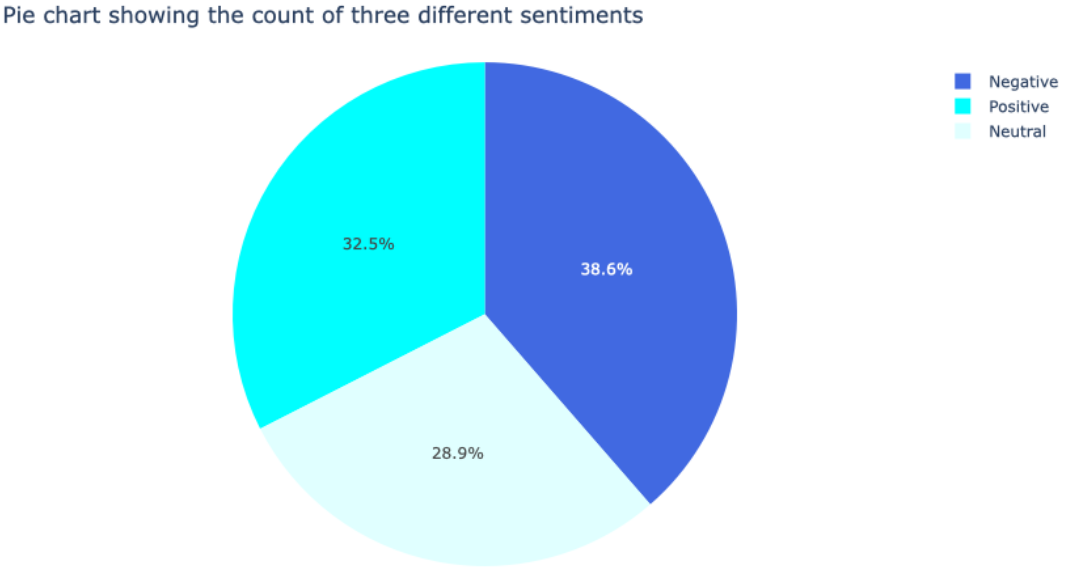


Figure 4.5: A pie chart showing three different sentiments: positive, negative, and neutral. The visual variables are color and the area of the chart segments.

The next task is the *extremum* task shown with a histogram plot in Figure 4.6. This plot has as visual variables color and the size of the bars (which is also related to the vertical position of the top border of each bar, as their bottom borders are aligned). Again the sum of the count of different tweets is visualized in different bins. Gray is used as standard bar color. The question to this chart is *What is the intensity score of the lowest sum of Count?*, with the options *0–0.05, 0.95–1, 0.5–0.55*, and *Don't know*. The correct answer is *0.95–1*.

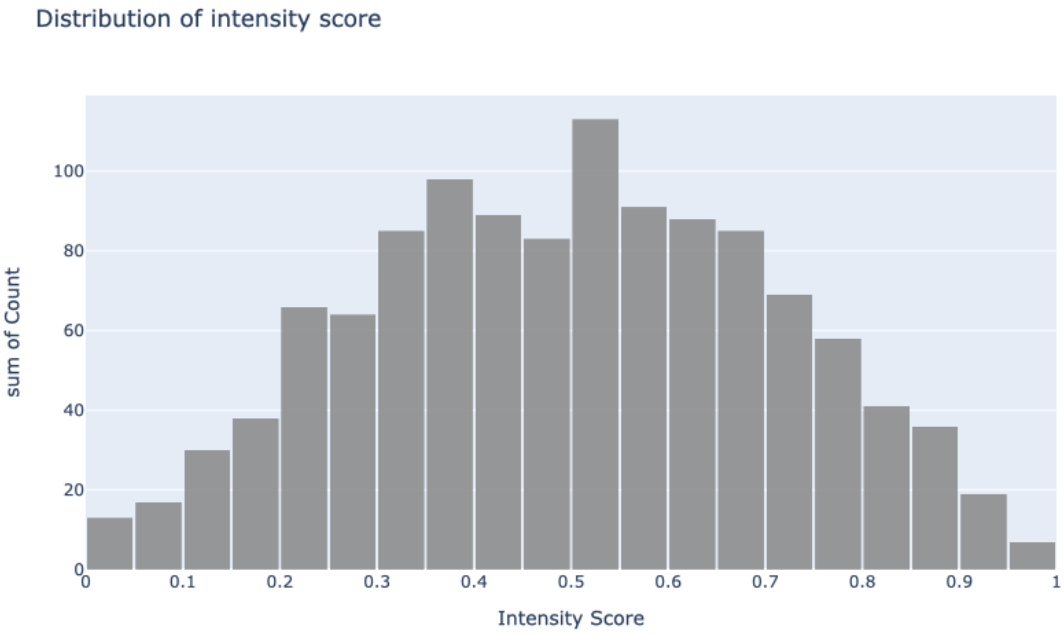


Figure 4.6: A histogram chart showing the intensity scores of different tweets. The visual variable involved for sentiment value encoding is the bar size.

The next and the last task is the *filter* task. Figure 4.7 shows a corresponding histogram plot to this task. The visual variables to this plot are color and the size of bars. The different sentiments are encoded as green for positive, red for neutral, and dark red for negative sentiment. The study asked the following question: *What rating range has between 6000 and 8000 total reviews?* with the options, *3–3.5, 3.4–3.7, 4–4.5*, and *Don't know*. The correct answer is *3.4–3.7*.

4.2 Study Design Implementation

This section explains the overall implementation procedure. All the visualizations have been implemented in Python using several visualization libraries. For most visualizations, *Plotly for Python*, *Bokeh*, and *Matplotlib* were used, and for the data handling, *pandas* was used. The primary challenge was managing the datasets, which varied in size. The general procedure was the following:

- Loading and trimming the datasets
- Choosing an appropriate visualization technique and selecting data to visualize
- Data handling and cleaning
- Creating visualizations and saving them as static images (a necessary step for interactive *Plotly* visualizations)

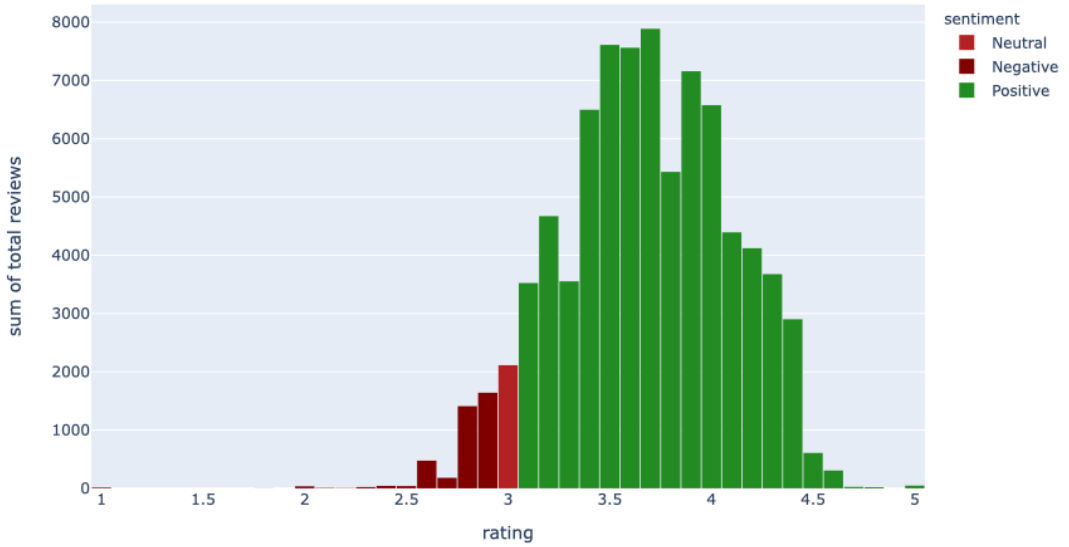


Figure 4.7: A histogram chart showing the intensity score for different tweets. The visual variables are the color and size of the bars. The chart is colored in three different ways: Red for neutral, dark red for negative, and green for positive sentiment.

The survey was implemented with an open-source JavaScript library *survey.js* and deployed as a *DigitalOcean* web application using *React.js*. This strategy was chosen over the more well-known alternatives (such as existing survey web portals) for the following reasons: (1) the survey creation framework should be highly customizable; (2) the questions should be displayed in a shuffled order; (3) the implementation should be able to track the time needed for answering a question. To the best of our knowledge, all these functionalities are not present in well-known free survey creation tools, such as *Google Forms*[1], *SurveyMonkey*[2], and more. Thus, a programming effort was necessary.

4.3 Further Considerations on the Experimental Design

To conclude the experimental design and the user study, it can be observed that the study is not exhaustive. There are many more sentiment visualization techniques available in the literature that are not included in this study. However, expanding the study further may lead to loss of generality. Our goal was to preserve a general audience and to perform the survey online. Expanding the questionnaire would increase the complexity and increases the user effort. Further, this study only focuses on a relatively small subset of predefined tasks. As already discussed in Section 2.5, sentiment visualization evaluation introduces several challenges. Thus, expanding the tasks set would also make the analysis of the results more complex.

In our work, we gathered as much information as possible to capture the expressiveness of our predefined usability measures. As previously mentioned, we use a set of low-level tasks by Amar et al. [16]. In Section 3.2.4 we already discussed that these tasks are well-known and widely used in related literature. We cover each task with different questions using different representation types. Thus, we convince our-

[1]Google Forms `https://www.google.com/forms/`
[2]SurveyMonkey `https://www.surveymonkey.com/`

selves that our user study is extensive enough to capture the visualizations' efficiency and effectiveness.

After conducting the study, we noticed that different questions are interpreted by users differently. Some questions also do not have a correct or wrong answer but are only used for preference analysis. This is discussed in much more detail in Section 6. To deal with subjective questions, we introduce several ranking, preference, and feedback questions on the last page of the survey. We gather information about the user's confidence in answering questions, visualization preferences, color preferences, and free form feedback. This helps us in overcoming the effects of unclear questions.

5 User Study Results

This section is concerned with reporting and initial analyses of the study results. It lists the results in an overview and shows the contribution to the research questions that this report should answer. Further, it visualizes and evaluates multiple aspects of the study results using different types of charts. We created all the following charts using the python library *seaborn* [76]. We start at the demographic information, followed by the actual survey questions, and lastly, the ranking and feedback questions. We also provide additional observations related to the conducted user study. Afterwards, we go into more detail about some interesting findings and further discussion in Section 6.

The online questionnaire was run between May 12 and June 6 for a total of 3.5 weeks. In the end, 50 participants took part in it. The raw data of responses and an aggregated list of response times, and the fraction of correct responses for each participant can be seen in Appendix B.

5.1 Demographic Information

Firstly, we show an overview of the participants who took part in the survey. For this, we list the results of the demographic questions on the first page of the survey. The four respective questions gathered the age, the gender, the highest degree or level of school completed, and if the person is currently studying. By analyzing the results, we got the following responses shown in Figure 5.1. The gender bar chart shows that participants were 24 females, 25 males, and 1 preferred not to say. The age bar chart shows the age distribution, where most participants are in the age category of 25–34 with 30 people, followed by 13 participants in the age group of 45–54, 3 in 55–64, 2 in 18–24 and 35–44. For the highest degree or level of school question, most participants hold a Master's degree, 19 participants to be exact. This response is closely followed by 14 participants holding a Bachelor's degree and 8 people with a doctorate. 4 people have a high school diploma or equivalent, 3 users have college credits but no degree, one person has a high school degree with no diploma, and another participant skipped the question. The next question is about if the user is studying. Most of the participants responded with *yes, as postgraduate / Ph.D. student*, 16 are studying but in another position that was not listed in the response options. 11 people responded *no*, and 3 participants voted *yes, as a university student*.

Next, we list the results of the remaining demographic questions without charts since they have a negligible effect on the further result analysis. Relevant results will later be discussed in detail. The next questions of the survey are *How would you rank your prior experience in creating visualizations?*; *Are you experiencing any color blindness issues?*; *On what device are you taking this survey?*; and *In what way are you familiar with creating visualizations?* Most of the participants answered they had limited experience, that is 18 users, followed closely by no experience with 14 participants, 10 with working knowledge, 6 with proficient knowledge, and 2 with expert knowledge. Most of the participants who had at least limited experience with visualizations were familiar with spreadsheet software and programming languages. Only one participant stated to have knowledge of Tableau software. The last two questions on color blindness and device type got the following responses. Most (47) of the participants had no color blindness issues. However, 3 participants stated to have minor color blindness issues. A total of 35 users took the study on a PC or a

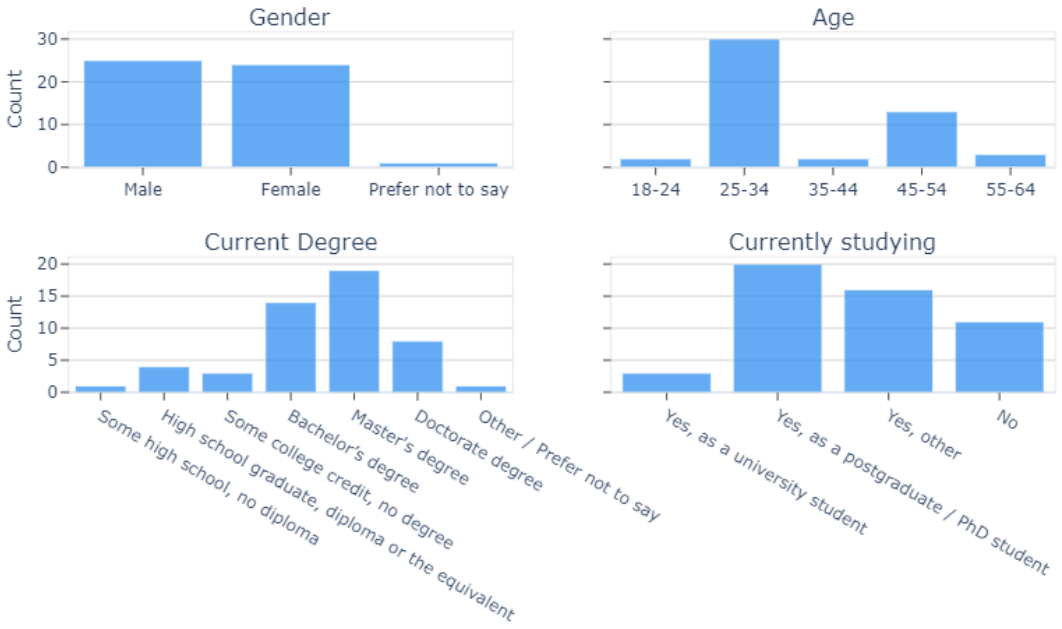


Figure 5.1: Response count of four different demographic questions from the user study.

tablet and 15 on their phone. It shows that a wide variety of users completed the survey. The mentioned aspects will be discussed in a later section.

5.2 Study Task Results

After showing the results of the demographic and warm-up questions, the following paragraphs take a deeper look at the user study's actual questions, tasks, and visualizations.

For an aggregated view of the raw results of the actual questions, we refer to Table B.1 in the Appendix B. The corresponding visualizations of these questions can also be found in Appendix A.

Some interesting observations can be drawn from the following two box plots. Figure 5.2 shows a box-plot of the average correctness rate distribution over each age group. Experience with visualizing software is depicted with color. Using this figure, several observations can be drawn. The analysis, however, strongly depends on the age distribution of participants. The group of people with age between 35–44 has the lowest mean correctness rate of around 73%. The highest correctness rate has the age group with people between the age of 55–64. However, this observation has to be taken with a grain of salt because the mean might change when more people take the study. Another interesting observation with the help of this plot is that the average mean between three different age groups is the same, with an average of around 87% success rate.

It is also to note that this box-plot and further figures used for results analysis and discussion only consider questions with a single correct answer. Some survey questions are ambiguous or have no correct answer. Thus, they are not included in these plots. This is discussed in more detail in Section 6.

The second box-plot in Figure 5.3 shows the average time per question. The age group of people between the age of 18–24 has the lowest average response time of around 18 seconds per question. It is followed by the age group 45–54 with an

average of about 23 seconds and closely followed by the age group 25–34 with an average of 23 seconds per question. Age groups with only a few (2–3) data points are not as trustworthy as others. From the age group, 18–24, one can see that participants with no prior visualization experience tend to be closer to the mean of response times than others.

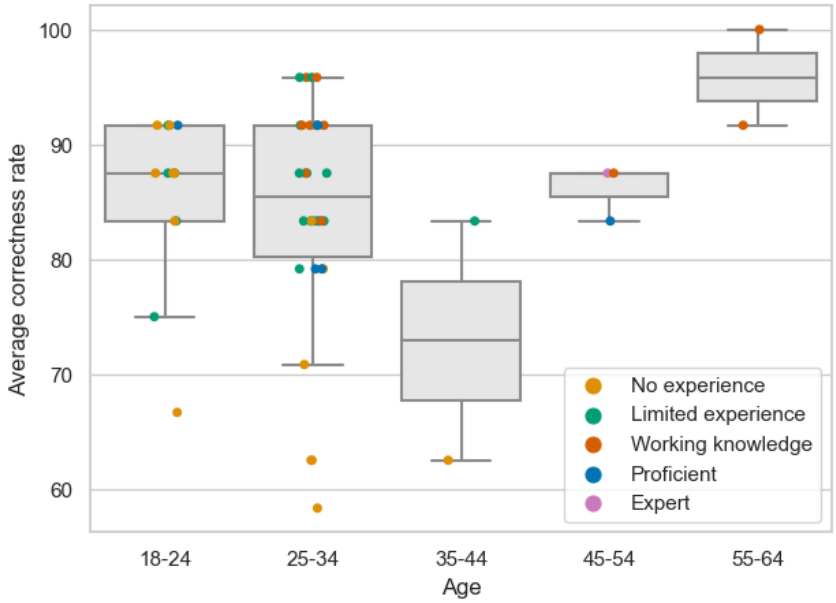


Figure 5.2: Distribution of the average correctness rate by age for all questions. The experience of creating visualizations is encoded using color.

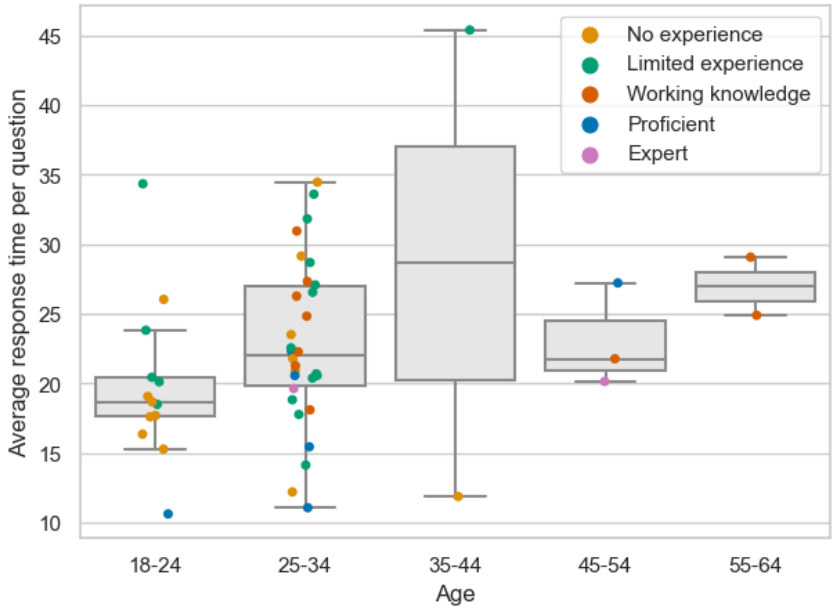
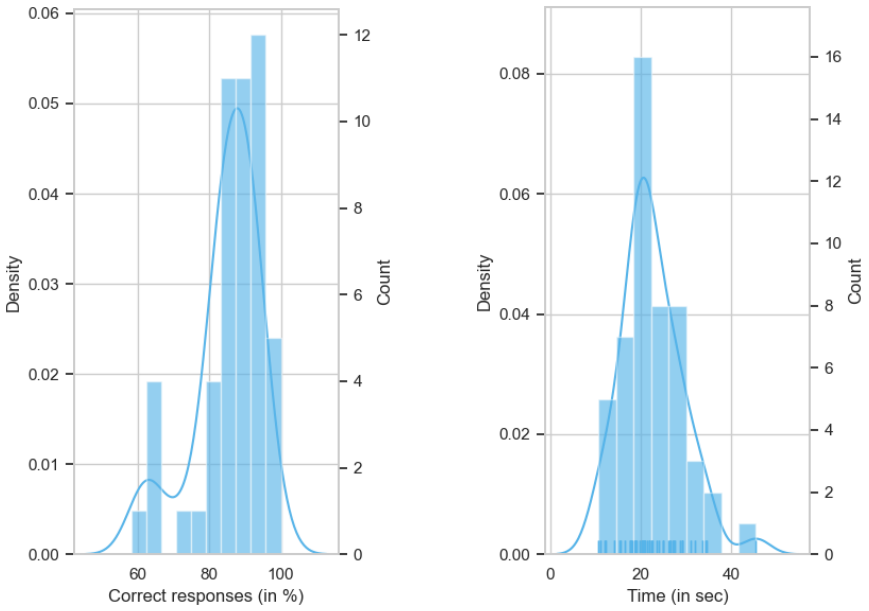


Figure 5.3: Distribution of the average response time by age per question. The experience of creating visualizations is encoded using color.

Figure 5.4 shows two histogram plots describing the distribution of correct responses. The left histogram plot shows the density and the right one the count of correct responses. As can be seen from the density plot in Figure 5.4(a), the majority

of participants answered around 90% of the questions right. The variance, that is, the spread of data points, lies between 85% and 100% of correct responses. The histogram with the counts can be used for a more detailed analysis of the density plot. Most people, that is more than 40 participants answered between 87% and 95% of the questions right.



(a) Correctness distribution (b) Response time distribution

Figure 5.4: (a) Average correct responses in percent with density and response count measures. (b) Average distribution of response times and response count.

Regarding the total count of the different response times captured in Figure 5.4(b), It can be seen that the average response time is around 20 seconds per question. The total average time spent per participant is around 10 minutes, which cannot directly be deduced from both plots. A total of 16 participants have response times that are close to the average.

Figure 5.5 shows a scatterplot visualizing the distribution of the average time per task with regard to the average error time. In this plot, the task name is encoded using color. The question with the lowest error rate and the lowest average time is a question from the *extremum* task. This question has an error rate of 0% which means that all participants answered this question correctly. The average response time of this question is around 12 seconds, which is the second-fastest answered question. From further analysis, the corresponding question to this data point is, *What phone brand has the lowest average rating?* using a *bar chart* (this information cannot be retrieved using the mentioned figure). Figure 5.5 also shows the visualization type used. With the use of these graphs, we can derive several observations. We can mainly group visualizations and tasks that perform better than others. We will take a more detailed look at this in the discussion section.

To analyze the confidence aspect of task solving, Figure 5.6 shows a plot with average response times per question vs. the correctness rate per user. Colors depict the different reported experience levels. From this graph, one can see that most of the participants had no, limited, or working experience levels in creating visualizations. Only a few stated to have proficient or expert knowledge. Results

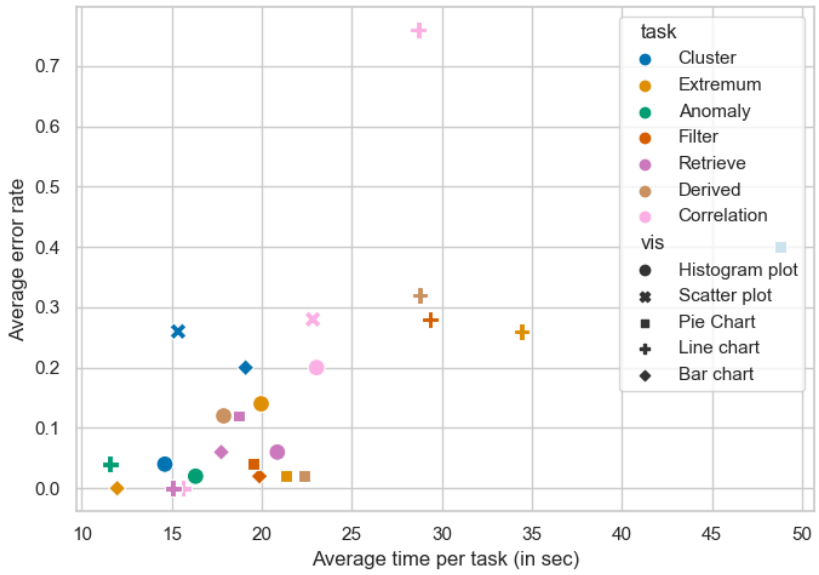


Figure 5.5: Distribution of average time per task vs. the average error rate per task. Where the task name is encoded using color and visualization type as symbols.

in the no experience group have the highest spread. The participants with working knowledge end up in one big cluster of points. This shows that people with some knowledge in visualizations tend to have a correctness rate of the average participant.

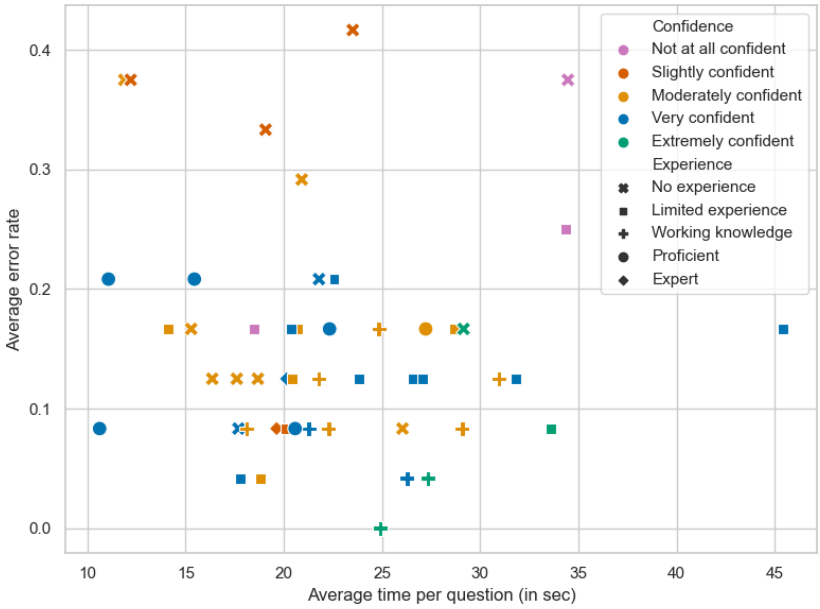


Figure 5.6: Average time per question vs. the error rate. Color visualizes user confidence in answering the questions, and symbols represent user experience in working with visualizations.

Figure 5.7 shows a histogram plot with multiple differently colored bars. The bars represent the user's experience of working with visualizations. It shows the count distribution with regard to the correctness response. In other words, it shows how well different users performed during the study. Bars on the right show users who

nearly answered all questions right. Each experience group has a curve that represents the distribution in the group. It is interesting that participants with limited experience or working experience tend to a normal distribution with a correct response rate of around 85%. The distribution of participants with no experience is unreliable as only two participants stated to have zero experience working with visualizations. Some participants have a correctness rate of 60–70% or 80–90%. For now, there is no interesting remark to be made about the distribution of participants with proficient or expert knowledge in visualization. We will go over this in more detail later.

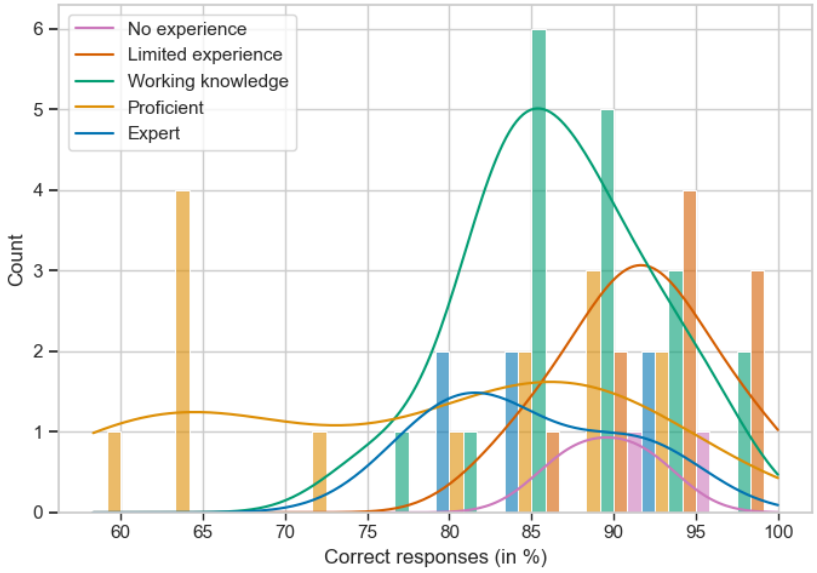


Figure 5.7: Distribution of correct responses by experience in working with visualizations.

In the first part of the questionnaire, we asked participants about their gender. We plotted a similar figure as the previous one and studied the average time per question vs. the average error rate ratio. We noticed that the gender of participants neither had an influence on response time nor the correctness rate. One main cluster could be identified in the range of 15–35 seconds for the response time and a correctness rate of 82–92%. This cluster contains nearly the same number of female and male participants. Some outliers could be detected for female and male participants.

Figure 5.8 shows a scatterplot with different task types vs. the average error rate per task. Color encodes the different visualization types. This graph can be used to derive tasks that fall out of the line, which means that the average error rate deviates a lot from other tasks. These tasks include only one task: a question from the *correlation* task using a line chart. The corresponding question to this is, *From the above graph, can you tell that there is a (weak) correlation between ratings and the number of reviews?* with the graph shown in Appendix A, Figure A.11. This may be due to the fact that this question was rather hard to be answered by the general audience. With the help of this diagram, one can further analyze different aspects of task and visualization types on the average error rate.

Figure 5.9 shows a scatterplot with different task types vs. the average response time per task. Color encodes the different visualization types. This graph can be used to analyze the average response times for different tasks and visualization types. The *clustering* task has a question with the highest average response time of around

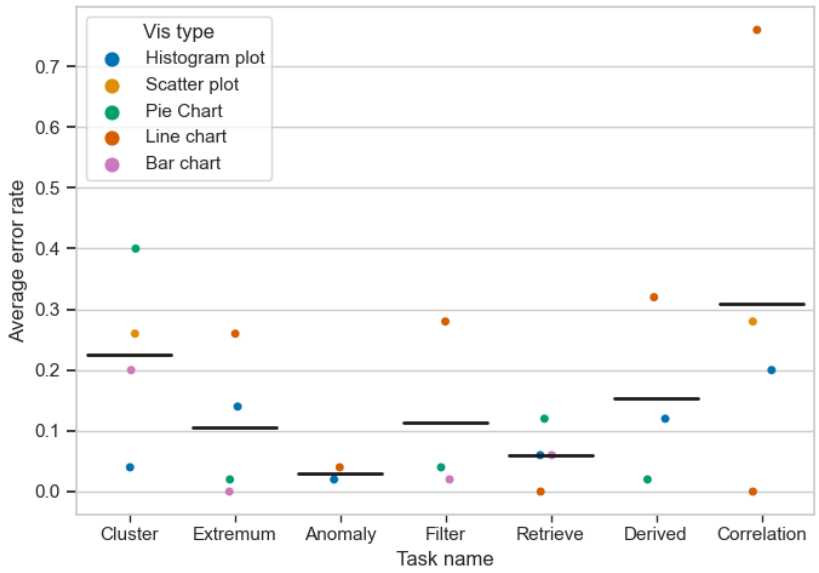


Figure 5.8: Distribution of different task types vs. the average error rate per task. Color encodes the different visualization types. The horizontal lines represent the mean of the average error rates per task.

48 seconds. The fastest response time is observed from the *anomaly* task using a line chart. The corresponding question has an average response time of around 12 seconds. Without the outlier of the *cluster* task, pie charts also tend to be around the average response time of 22 seconds of all tasks. The same observation can be drawn for histogram plots. Other types of visualization yield unclear results.

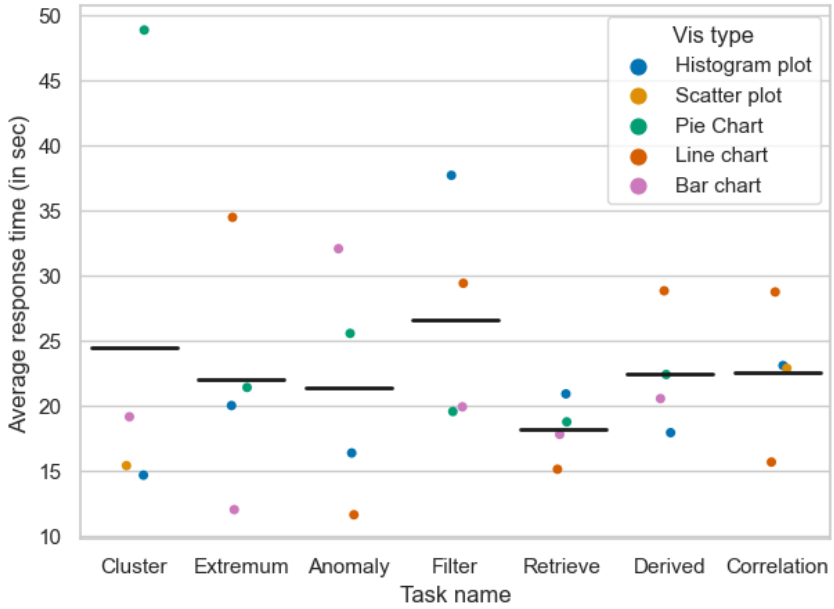


Figure 5.9: Average response time per task with visualization type encoded in color. The horizontal lines represent the mean of the average response times per task.

We further studied if the chosen device type of users had an influence on the overall study performance. By plotting similar plots as the ones discussed, we made some interesting observations. It can be seen that different users performed much

worse in regard to the global average. These outliers, however, don't depend on the device type. It can be seen that participants taking the survey on the phone get around the same average response time and correctness rate as people taking it on their PC or tablet. This clearly shows that neither response time nor correctness rate depends on the device used. This, however, only holds for these specific user tasks, visualization types, and datasets. These and further observations can be drawn from Figure 5.10.

Figure 5.10 shows a scatterplot visualizing the average time per question over the average error rate with which participants solved each question. Color shows if participants stated in the demographic questions section if they are currently studying. Analyzing this figure, we notice a rather equal distribution of general, postgraduate/Ph.D. students, and others. This strengthens the fact that this survey was meant to be answered by the general public. Even one entry shows a participant with an error rate of 0%, despite saying that they were not currently studying.

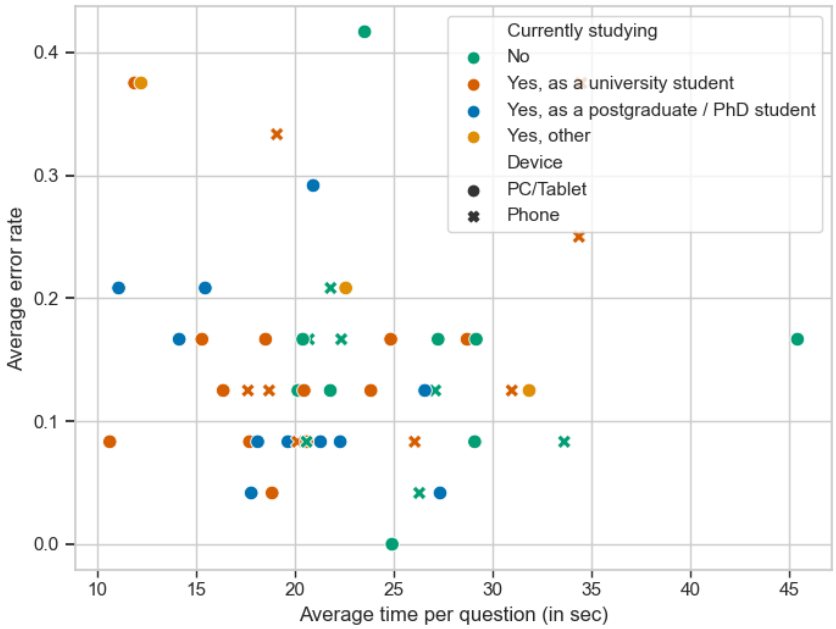


Figure 5.10: Average time per question vs. average error rate with currently studying encoded in color and device type using symbols.

5.3 User Preferences and Further Feedback

The next part is the ranking and feedback questions from the last page of the study. It consists of the following close-ended questions. *How confident were you in answering the previous questions? Overall, what visual representation did you find most easy to work with? (Easy to hard ranking), Some plots used different colors for different sentiments and emotions. Did you find some colors easier to work with than others? What color (in general) would you find best for identifying positive and negative emotions?* and a form for users to leave feedback and comments about the survey. Most participants were moderately or very confident in answering the questions, with 38 participants selecting one of both options. Followed by 5 participants who stated that they are slightly confident about answering the questions. 4 users were extremely confident, and 3 users were not confident at all in answering the questions.

Regarding the preference question about the visualizations, 36 participants stated that some colors were easier to work with than others. 10 stated that they did not perceive a difference in efficiency, and 3 stated they do not know. One participant skipped this question. Most (46) of the users stated that their color preference is *Green for positive and red for negative sentiment*, 4 stated that *white for positive and black for negative* is their preference. 4 users stated other preferences, and 2 of them wrote: *"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."*. It is to note that these results do not add up to the total amount of participants (50) since this question was a multiple choice type question. Thus users had the chance to select more than one preference. Table 5.1 shows the top 10 ranking preference of different visualizations.

Table 5.1: Top 10 chart types sorted by users' preference. Sorted from left to right and top to bottom.

Ranking order (left to right and top to bottom)	Response count
[Bar chart, Histogram chart, Line chart, Pie chart]	4
[Pie chart, Histogram chart, Bar chart, Line chart]	3
[Histogram chart, Bar chart, Pie chart, Line chart]	3
[Line chart, Pie chart, Bar chart, Histogram chart]	3
[Histogram chart, Bar chart, Line chart, Pie chart]	2
[Pie chart, Bar chart, Line chart, Histogram chart]	2
[Line chart, Bar chart, Histogram chart, Pie chart]	2
[Histogram chart, Line chart, Pie chart, Bar chart]	2
[Bar chart, Line chart, Histogram chart, Pie chart]	2
[Bar chart, Pie chart, Histogram chart, Line chart]	2

Another question was in the form of a text input area where users could give general feedback about the study. We received the following comments:

- *"It's odd to ask for a total amount given in percentage... Question 30 and the options of answer are confusing. "No" and "Don't know" could be interpreted as the same."*
- *"The graphics are, in my opinion, a bit too small."*
- *"Didn't understand the usage of question 41 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"."*
- *"I enjoyed that very much. Very interesting. Good luck!"*
- *"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."

5.4 General Observations

Finally, we can share some general observations and further insights into the data. First, we consider the feedback form, which was filled out by six participants. One participant noted that pie charts are generally not that practicable to work with, which we consider as well in the discussion part. One participant complained about the usability of the ranking question on the last page of the study. It is to note that this user took the survey on their phone. This states that this question is not reliable when answering it on mobile devices. Nonetheless, in our experiments, we tested the survey on multiple portable devices. This showed no issues, and the widget was working fine, and we could not reproduce the problem mentioned by the user. The results also show that most people who took the survey on their phones were, in fact, able to interact with the ranking widget. Another participant mentioned that it would be odd to ask a total amount given in a percentage value. This would be confusing in the question formulation, but most of the participants still answered the concerned question correctly. Another user stated that they enjoyed the study and found it interesting. Another user stated that they misclicked the wrong answer and could not change their option. This was considered in the result analysis and filtered out from the aggregated overview. As a side note, due to the survey's time logging feature, this behavior is anticipated. Otherwise, this feature could not be implemented reliably.

6 Discussion

This section further discusses the results and shows how it answers the research questions of this work. For this, we focus on the different research questions with a look at the study's results. We discuss the results on the effectiveness and efficiency using the definitions by Frøkjær et al. [28]. In addition, we present multiple guidelines for researchers and practitioners that should help in decision-making. We divide the discussion into several parts to cover each aspect of the study and the research questions. As already seen in this report's introduction and background section, these terms are widely used in visualization evaluation. The discussion of these metrics gives a good base for the usability of different visual representations and metaphors. It is critical to take a deeper look at the analysis of the prior results, as these metrics highly depend on various dependent variables.

6.1 On Effectiveness

Indicators of effectiveness include quality of solution and error rates. In this study, we use both metrics as the primary indicator of effectiveness, which measures the outcome of the user's interaction with the system. Due to space limitations, we present a table with an aggregated view of the received answers in Appendix B. It shows an overview of all questions with the count of correct and wrong answers. As this table contains all of the acquired information of the study, we can use it to start the discussion. The overall performance was pretty decent as most of the users had error rates close to the global average. Only a few users had above-average error rates.

The following two questions were answered incorrectly by most of the participants. *How many positive sentiments can you identify in the above graph?* and *From the above graph, can you tell that there is a (weak) correlation between ratings and the number of reviews?*. From the first one, it can be derived that people interpret various emotions in different ways. This question asks the participants how many positive sentiments they could identify in the corresponding graph. This also includes the sentiments/emotions of *surprise* and *trust*. It is worth noting that these two sentiments are not well-defined in polarity, and it is indeed up to the user's preference if these should be considered negative, positive, or neutral. The second question is concerned with the presence of a weak correlation in the corresponding visualization. The ratio of responses is 12 (Don't know) to 12 (Correct answers) to 26 (Wrong answers). Most participants stated that they see a weak correlation in the line plot. This is due to the fact that the question might not be clearly defined. At some timesteps, there is indeed a correlation between ratings and the number of views. However, this correlation is only local and not global. As these are the two topmost questions with controversial responses, the remaining questions only show a weak correct to wrong answer ratio. For details, we refer back to the table in Appendix B. It shows a ranking from the most correct to the most incorrectly answered questions.

In the next paragraphs, we discuss the effectiveness in the context of different tasks, visualization types, and visual variables. For this, we highlight relevant questions in a few figures. To begin with, we take a look at a few different tasks. As seen in Figure 5.8 (as well as the legend of Figure 6.1), the study consisted of seven different tasks.

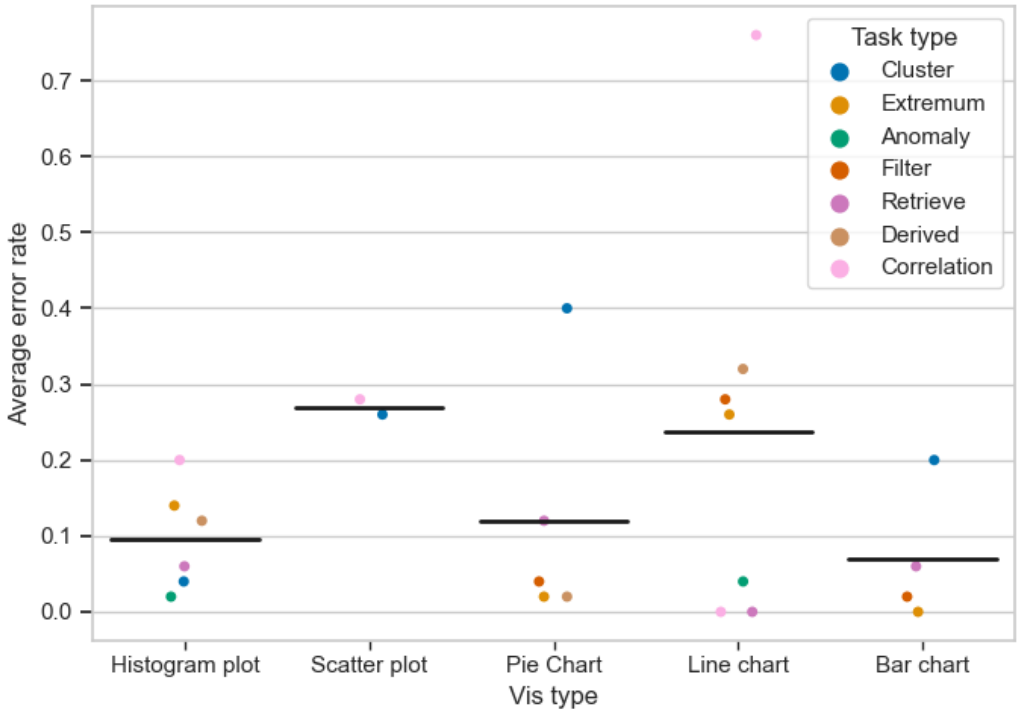


Figure 6.1: Distribution of the average error rate per question categorized by the visualization type. The horizontal lines represent the mean of the average error rates per representation type.

Participants were the most effective in the following tasks, sorted from most to least effective: *retrieve*, *extremum*, *filter*, *derived*, *cluster*, and *correlation*. The *anomaly detection* task is left out because it consisted only of two questions, so it is unreliable enough to evaluate its effectiveness. This is calculated with the average error rate over all questions of a specific task for all participants. The means are highlighted in Figure 5.8. Switching from the task to representation perspective, we can note that the users were the most effective in the following type of visualizations, sorted from most to least effective: bar chart, histogram plot, pie chart, line chart, and scatter plot. These conclusions can be drawn from Figure 6.1. It is to note that the survey questions only included two scatter plots. Thus the observation of the effectiveness of these plots might not be reliable.

Regarding visual variables, the list of preferences consists of the following channels in descending order: color, bar size, and point's position. For marks, the ordering is as follows: points, lines, and then area. These insights can be drawn from Table 5.1 of user's preference of working with different chart types. Four questions had no correct or wrong answers but were rather used for preference evaluation. These questions are written in italic in the raw results table found in Appendix B. From the results of these questions, the following insights can be drawn on sentiment color preference. Participants identified the polarity of the sentiments/emotions of *surprise* and *trust* differently. Only 11 participants stated these two sentiments as positive polarity, 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. A similar conclusion can be drawn from another question of this set.

6.2 On Efficiency

Efficiency can be measured in different ways, for example by calculating task completion time and learning time [28]. As our study is not designed for measuring learning time, we use task completion time to measure efficiency. For this, the response time for each question is tracked and logged.

We start by giving an overview of the most efficient tasks and visualization types. The previously introduced Figure 5.4(b) shows some statistical details about the total times of all participants. It shows the average and standard deviation of response times. Using Figure 5.9, it is already clear which tasks seem to be simpler, thus more efficient than others. The following insights can be drawn from this figure. The simpler tasks regarding the average response time in ascending order are *retrieve, anomaly*, *extremum*, *derived*, *correlation*, *cluster*, and *filter*. With a similar approach, we get the most efficient visualization types from Figure 6.2 in the following order, from most to least efficient: scatter plot, bar chart, histogram plot, line chart, and pie chart.

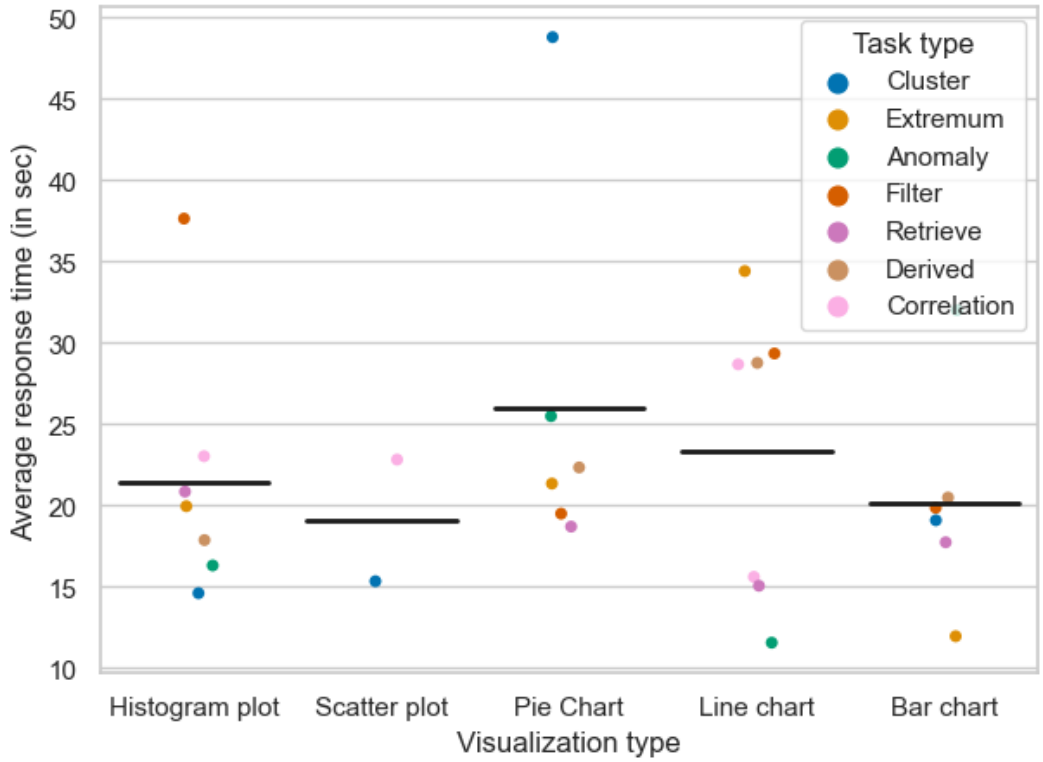


Figure 6.2: Distribution of the average response time per question categorized by the visualization type. The horizontal lines represent the mean of the average error rates per representation type.

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. It is closely followed by the following questions with an average response time between 31 and 37 seconds:

What rating range has between 6000 *and* 8000 *total reviews?* using a histogram plot, *The above graph shows a distribution of sentiment intensity scores. At what point does the data look abnormal?* using a histogram plot, and *At what time window was the day with the maximum number of negative tweets?* using a line plot. These and further findings can be drawn from Figure 6.2. To not overload this section with figures and tables, the corresponding questions can only be identified by manually looking at the raw results of the study, which again can be seen in Table B.1 in Appendix B. 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 couldn't identify differences in efficiency. However, many participants noticed graphs where polarity seemed to have a "wrong" color. This means that they 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 less and more efficient visual channels. We thus come to the same conclusion as before with regard to channels, ordered from most to least efficient: color, bar size, and point's position. And for marks, the ordering is again: points, lines, and area.

6.3 Other Results

As our survey also showed some interesting results on user satisfaction and free-form user feedback, we also take some short notes on these results. We take the definition for satisfaction from Frøkjær et al. [28], which is the users' comfort with the use of the system. In this setting, "the system" corresponds to the contents of the user study.

We use the questions on the study's last page to determine which visualization types people prefer. Most of the participants stated that they were moderately or very confident in answering the questionnaire. 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. We conclude that it is best to choose these three colors, or at least green and red, for positive and negative sentiments.

In the previous section, we listed the feedback we got from the participants. Overall the feedback was very positive, and only a few statements were made about some confusing design choices. We highly appreciated this feedback. Thus we also considered this in our discussion.

6.4 Guidelines

This subsection lists several guidelines for practitioners who need to use basic sentiment visualization techniques for their work. We came up with these guidelines after the analysis of the study results. We list guidelines about the different types of user tasks used in this work. We further list guidelines regarding usability measures that answer the research questions in multiple ways. Additionally, we give a global overview of guidelines.

We start by identifying multiple task-based guidelines, which are a starting point when choosing task-specific visualizations.

- *Anomaly detection*: Both histogram plots and line plots showed a low average error rate for the *anomaly detection* task. However, the accuracy was higher on

line plots, and they had an overall low response time. Therefore we recommend using line plots over histogram plots for suitable data.

- *Cluster identification*: Only scatter plots outperformed all the others regarding response time and accuracy. We recommend using scatter plots over pie charts since pie charts showed high response times and slightly below average error rates. Histogram plots and bar charts showed average performance. Thus we recommend using them only when necessary.

- *Correlation detection*: The *correlation detection* task had the highest error rate of all tasks. Therefore it is critical to select appropriate visualizations. We recommend using line plots as this was the only visual representation with an error rate of 0%. Histogram plots and scatter plots scored average performance. Pie charts and bar charts were not evaluated regarding this task. Thus we don't recommend using them as it is hard to use these types of visualizations to visualize correlations.

- *Compute a derived value*: For *computing a derived value*, we recommend using pie charts as these had the lowest average error rate among others. They were followed by histogram/bar charts, line charts, and scatterplots. Overall these representations showed similar average response times for this task.

- *Find extremum*: We recommend using bar charts for finding *extremum* values. This is not surprising as it is easy to identify maximum and minimum values in a bar chart quickly. Bar charts had an error rate of 0% and the lowest response time average of around 12 seconds. Following bar charts, we recommend using pie charts, histogram plots, and line charts in the respective order. Line charts are recommended here as the last option since these can get very confusing when visualizing multiple lines. However, retrieving *extremum* values on other charts is easy, no matter how many sentiments are visualized.

- *Filter data*: We recommend using bar charts and pie charts for the *filter data* task. In our study, each of these visual representations performed equally well. Both representations average on an error rate of around 4–5%, which is negligible. They have average response times close to the global average. We do not recommend using line charts for this type of task as it is quite hard and time-consuming filtering exact values from this type of plot.

- *Retrieve value*: To master the *retrieve value* task, all visualization types showed above-average performance. For this, we recommend using any visual representation from the group of pie charts, bar charts, scatter/line plots, and histogram plots.

In the previous paragraph, we listed different guidelines regarding different user tasks. Now we introduce several guidelines by identifying visualization techniques and metaphors that perform well with regard to usability definitions.

- *On effectiveness*: Regarding the effectiveness of different visualization representations, we recommend using plots in the following order: bar charts, histogram plots, pie charts, and then line and scatter plots. These types of charts performed overall well in our study. Overall, pie charts should be avoided as they often lead to misinterpreting sentiment polarity, thus leading to low effectiveness.

- *On efficiency*: Regarding the efficiency of different visualization representations, we recommend using plots in the following order: bar charts over histogram plots over line/scatter plots over pie charts. Again, pie charts show the lowest efficiency. This is again as expected, as these chart types only allow to efficiently represent a small number of sentiments.

- *On user satisfaction*: Concerning user satisfaction, we take another look at how participants perceived different visualization types and metaphors. Already in Section 5 we mentioned with the help of follow-up questions which visualizations earned the most positive feedback. However, we do not see a clear separation of user satisfaction between different visualization methods. Therefore, it strongly depends on the type of users of the experimental setting and what main task is to be achieved. Most users stated that they liked using bar charts, histogram plots, and pie charts over line and scatter plots.

We now discuss how visual variables/channels or marks should be selected. For best practices, we recommend selecting color as the visual variable to represent sentiments. Most of the participants voted for green for positive and red for negative sentiments. However, some users also stated to keep the same scheme and a consistent use of visual channels in different visualizations. For example, color should always represent the same sentiments in different visualizations. Another interesting observation is that some users mentioned using a combination of white and black or green and black for sentiment polarity, more specifically, white and green representing positive sentiment polarity, and black—negative polarity. This was, however, suggested only by a minority of the users. Therefore we rank this guideline lower than the previous one, and this should only be used if the general setting of the visualizations is grayscale. To avoid distracting readers from multiple different colors, this path should be chosen.

As a general approach, we recommend following the ranking of visualization types from most to least effective/efficient based on our results in the context of sentiment visualization: *bar chart* > *histogram chart* > *line chart* > *pie charts*.

6.5 On Visual Variables, Channels, and Marks

To recall, the research question **RQ1** stated in Section 1.4 is: *Which visual variables/channels are most effective and efficient with regard to visual encoding of polarity and emotions?* This gets addressed by the comparative evaluation of several encodings using alternative visual variables. For example, color for emotions; or the height of bar charts for rating scores. After analyzing the results on the effectiveness and efficiency of multiple aspects of sentiment visualization of the previous sections, the research question is answered. We saw that color, bar size, and a point's position are the most important visual channels in terms of effectiveness. The most effective visual marks are points, lines, and area. For efficiency, we observed similar patterns. Users are more efficient in identifying sentiments when the right colors are used. For this, people tend to perceive positive sentiments for green and negative sentiments for red color. This was observed by analyzing the results on different aspects. Many characteristics of visual variables may influence the effectiveness and efficiency of the visual encoding of sentiments. Thus, it is critical to follow different guidelines.

6.6 On Visual Metaphors and Representations

The research question **RQ2** is: *Which visual metaphors/representations are most effective and efficient with regard to visual encoding of aggregated sentiment or opinion data?* This gets addressed by the comparative evaluation of several encodings using alternative visual metaphors. For example, using different chart types. Each chart type has its different advantages and drawbacks. After analyzing the results on the effectiveness and efficiency of multiple aspects of sentiment visualization, the research question is answered. These usability metrics are based on a lot more than just a few different visualization types. They depend on, for example, the type of task to be achieved, what the target audience is, and what datasets are used. However, our study showed that several representation types performed better than others. Therefore we listed several representation types that fulfill different usability measures. Concerning efficiency, we came up with the following ordering from most to least efficient, scatter plots, bar charts, histogram plots, line charts, and pie charts. We noticed that participants were the most effective with bar charts, histogram plots, pie charts, and then line and scatter plots.

6.7 Color Perception Issues

Three user study participants mentioned that they experience minor color blindness issues. From our analysis, we did not discover any disadvantage with regard to participants with no color perception issues. It is, however, essential to reflect on this issue further. Zhou and Hansen [77] particularly take a look into different colormaps. They conduct a survey to discover relevant colormaps, which may assist readers and researchers in choosing the right colormap for their own applications. Harrower and Brewer [78] implement a visualization tool for choosing effective color schemes for thematic maps. Their work fills existing gaps in the usability of graphic software by providing a tool to identify how to use color schemes most effectively. They also consider the option to choose colorblind-friendly colors. Also in this report, we used such a color scheme. In Section 5 we generated figures only using colorblind-friendly colors. We used the built-in feature from the Python library *seaborn* [76] to select an appropriate color scheme. Figure 6.3 shows six different colormaps already implemented in *seaborn* [79], including colorblind-friendly colors.

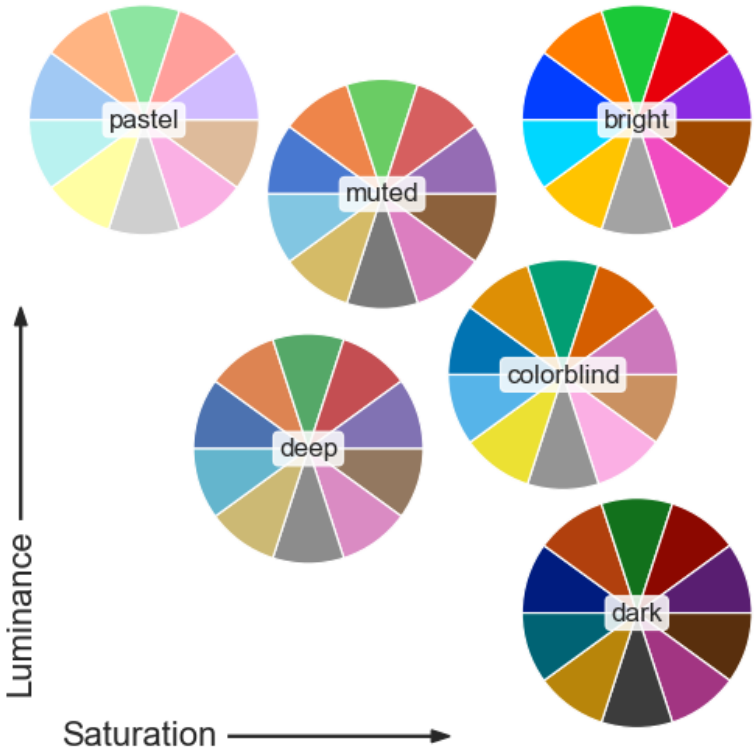


Figure 6.3: Six different colormaps available in *seaborn* including a colorblind-friendly colormap [79].

7 Conclusions and Future Work

In the final section of this thesis report, we provide a short summary of the project results and the answers to the research questions formulated at the beginning of the report. Afterwards, we discuss several directions for possible future work related to the topic and the findings of this work.

7.1 Conclusion

There is a vast amount of sentiment visualization techniques available in related literature. However, these often lack in evaluations about different usability measures such as effectiveness, efficiency, performance, and user satisfaction. This work addressed this issue and looked at basic high-level visualizations in terms of these different usability measures. We conducted a task-based user study with participants from a general audience. We designed several sentiment visualization figures by using basic chart types known from general information visualization. The survey consisted of seven tasks and 28 well-defined questions. After the analysis of the results, we discussed how different visual variables and representations influence the mentioned usability measures. We came up with many different guidelines concerning different user tasks and visualizations. Researchers and practitioners can then use these guidelines as a first starting point to develop high-quality and reliable sentiment visualizations that support usability.

7.1.1 Answers to the Research Questions

The research questions of this work formulated in Section 1.4 guided us through this research to address usability in different ways.

- **RQ1**: Which visual variables/channels are most effective and efficient with regard to visual encoding of polarity and emotions?
- **RQ2**: Which visual metaphors/representations are most effective and efficient with regard to visual encoding of aggregated sentiment or opinion data?

The following summarized guidelines answer the first research question. Select color as a visual variable to represent sentiments or emotions. Most of the users of the study stated that green should be used for positive sentiments and red for negative sentiments. For best practices, keep the same color scheme between different visualizations. In a grayscale setting, white should be used for positive and black for negative emotions. The second research question is answered by the following summarized guidelines. It is hard to fix one single representation for the most effective and efficient visualization. In general, use bar charts over histogram charts over line charts over pie charts. For task-based approaches, we refer back to the previous section, where we discussed particular task-based guidelines.

7.2 Future Work

The experimental findings should be viewed in light of the visualization tasks and datasets that were defined. Common visualization techniques from related literature were tested. However, the results should be viewed in the sense of the given settings. Nonetheless, more research is needed to test the research questions by using various

visualization tools, tasks, and datasets. Participants in this study were asked to complete tasks using static visualizations. While we realize the value of interactivity and how it can affect the user's experience with visualization [2, 15, 34], we decided to leave it out for the following reasons. First, including interactivity increases the study's complexity. In reality, one would have to consider a different set of variables, for instance, the users' input devices, such as a mouse, touchscreen, etc. Also, interaction design and implementation had to be considered. Each interaction is implemented differently on different input devices. Static visualizations, on the other hand, are often used for presentations and educational purposes. For example, visualization used in books, newspapers, presentations, etc. In all of these instances, representations are required to use static visualizations to complete several tasks. For this, we encourage more research into the effectiveness and efficiency of these visualizations while taking interactivity into account.

The number of visual marks shown in the visualizations used in this work is limited to a small number due to the practical limitations of performing the study using static visualizations with a large number of visual marks. For example, the length and complexity of the experiment. However, we would like to point out that the output of these visualizations can vary depending on the number of data points encoded. The findings of the study apply to static visualizations with a few visual marks and low complexity. Future research may look at how data point cardinality affects the task-based quality of visualizations. The effectiveness and efficiency of four basic two-dimensional visualization forms were investigated in this work. Some visualization forms, however, may be expanded to more than two dimensions. Depending on their dimensionalities, the performance of these visualization styles can vary. Investigating the influence of the number of dimensions expressed by a visualization form on its effectiveness and efficiency is an important line of research to pursue.

Finally, we should remember the concern discussed by Isenberg et al. [30] about the evaluation of larger complex visualization / visual analysis approaches, which cannot be reduced to controlled experiments focusing on individual visual representations and low-level interactions. This challenge is still open in regard to sentiment visualization problems—thus, further evaluation of such larger 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 important future research.

References

[1] J.-D. Fekete, J. J. van Wijk, J. T. Stasko, and C. North, "The Value of Information Visualization", in *Information Visualization: Human-Centered Issues and Perspectives*, ser. LNCS, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds., vol. 4950, Springer, 2008, pp. 1–18.

[2] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[3] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual Analytics: Definition, Process, and Challenges", in *Information Visualization: Human-Centered Issues and Perspectives*, ser. Lecture Notes in Computer Science, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds., Springer, 2008, pp. 154–175.

[4] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010.

[5] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise Data Analysis and Visualization: An Interview Study", *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2917–2926, Dec. 2012.

[6] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim, "Visual Analytics for the Big Data Era — A Comparative Review of State-of-the-art Commercial Systems", in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, ser. VAST '12, IEEE, 2012, pp. 173–182.

[7] R. C. Roberts and R. S. Laramee, "Visualising Business Data: A Survey", *Information-an International Interdisciplinary Journal*, vol. 9, no. 11, 2018.

[8] M. Behrisch, D. Streeb, F. Stoffel, D. Seebacher, B. Matejek, S. H. Weber, S. Mittelstädt, H. Pfister, and D. Keim, "Commercial Visual Analytics Systems — Advances in the Big Data Analytics Field", *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 10, pp. 3011–3031, Oct. 2019.

[9] J. Heer, M. Bostock, and V. Ogievetsky, "A Tour through the Visualization Zoo: A Survey of Powerful Visualization Techniques, from the Obvious to the Obscure", *Queue*, vol. 8, no. 5, pp. 20–30, May 2010.

[10] J. Heer and B. Shneiderman, "Interactive Dynamics for Visual Analysis: A Taxonomy of Tools That Support the Fluent and Flexible Use of Visualizations", *Queue*, vol. 10, no. 2, pp. 30–55, Feb. 2012.

[11] H. Mei, Y. Ma, Y. Wei, and W. Chen, "The Design Space of Construction Tools for Information Visualization: A Survey", *Journal of Visual Languages & Computing*, vol. 44, pp. 120–132, Feb. 2018.

[12] M. S. T. Carpendale, "Considering Visual Variables as a Basis for Information Visualisation", PRISM / University of Calgary, 2001-693-16, Jan. 2003.

[13] C. Ware, *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., 2004.

[14] O. Kulyk, R. Kosara, J. Urquiza, and I. Wassink, "Human-Centered Aspects", in *Human-Centered Visualization Environments: GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany, March 5-8, 2006, Revised Lectures*, ser. LNCS, vol. 4417, Springer, 2007, pp. 13–75.

[15] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", in *Proceedings of the IEEE Symposium on Visual Languages*, ser. VL '96, 1996, pp. 336–343.

[16] R. Amar, J. Eagan, and J. Stasko, "Low-Level Components of Analytic Activity in Information Visualization", in *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, ser. INFOVIS '05, IEEE Computer Society, Oct. 2005, p. 15.

[17] C. Görg, M. Pohl, E. Qeli, and K. Xu, "Visual Representations", in *Human-Centered Visualization Environments: GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany, March 5-8, 2006, Revised Lectures*, ser. LNCS, A. Kerren, A. Ebert, and J. Meyer, Eds., vol. 4417, Springer, 2007, pp. 163–230.

[18] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text", *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, Apr. 2014.

[19] D. Tanna, M. Dudhane, A. Sardar, K. Deshpande, and N. Deshmukh, "Sentiment Analysis on Social Media for Emotion Classification", in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 911–915.

[20] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[21] C. C. Aggarwal, *Machine Learning for Text*. Springer, 2018.

[22] B. Pang and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[23] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, Mar. 2013.

[24] K. Ravi and V. Ravi, "A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications", *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[25] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1253, 2018.

[26] S. M. Mohammad, "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text", in *Emotion Measurement*, H. L. Meiselman, Ed., Woodhead Publishing, 2016, pp. 201–237.

[27] K. Kucher, C. Paradis, and A. Kerren, "The State of the Art in Sentiment Visualization", *Computer Graphics Forum*, vol. 37, no. 1, pp. 71–96, Feb. 2018.

[28] E. Frøkjær, M. Hertzum, and K. Hornbæk, "Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '00, ACM, 2000, pp. 345–352.

[29] A. Shamim, V. Balakrishnan, and M. Tahir, "Evaluation of Opinion Visualization Techniques", *Information Visualization*, vol. 14, no. 4, pp. 339–358, 2014.

[30] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A Systematic Review on the Practice of Evaluating Visualization", *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2818–2827, Dec. 2013.

[31] D. Skau and R. Kosara, "Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts", *Computer Graphics Forum*, vol. 35, no. 3, pp. 121–130, Jun. 2016.

[32] H. C. Purchase, *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press, 2012.

[33] N. Elmqvist and J. S. Yi, "Patterns for Visualization Evaluation", in *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, ser. Information Visualization, vol. 14, Jul. 2015, pp. 250–269.

[34] T. Munzner, *Visualization Analysis & Design*, ser. A.K. Peters visualization series. CRC Press Taylor & Francis Group, 2014.

[35] L. Byron and M. Wattenberg, "Stacked Graphs — Geometry & Aesthetics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1245–1252, Nov. 2008.

[36] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets", in *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, 2018.

[37] SemEval-2018. (2018). "SemEval-2018 Task 1: Affect in Tweets (AIT-2018)", [Online]. Available: `https://competitions.codalab.org/competitions/17751` (visited on Sep. 5, 2021).

[38] G. Nibras. (2019). "(Un)Locked Cell Phone Ratings and Reviews on Amazon", [Online]. Available: `https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews/version/1` (visited on Sep. 5, 2021).

[39] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. L. Arendt, S. Shaikh, and W. Dou, "Vulnerable to Misinformation?: Verifi!", in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ACM, Mar. 2019, pp. 312–323.

[40] T. Kulahcioglu and G. de Melo, "Paralinguistic Recommendations for Affective Word Clouds", in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, ACM, 2019, pp. 132–143.

[41] K. Watson, S. S. Sohn, S. Schriber, M. Gross, C. M. Muniz, and M. Kapadia, "StoryPrint: An Interactive Visualization of Stories", in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, ACM, 2019, pp. 303–311.

[42] J. Chamberlain, U. Kruschwitz, and O. Hoeber, “Scalable Visualisation of Sentiment and Stance”, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, ser. LREC 2018, European Language Resources Association (ELRA), 2018, pp. 4181–4182.

[43] E. Cuenca, A. Sallaberry, F. Y. Wang, and P. Poncelet, “MultiStream: A Multiresolution Streamgraph Approach to Explore Hierarchical Time Series”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 12, pp. 3160–3173, Dec. 2018.

[44] S. Fu, Y. Wang, Y. Yang, Q. Bi, F. Guo, and H. Qu, “VisForum: A Visual Analysis System for Exploring User Groups in Online Forums”, *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 1, 3:1–3:21, Feb. 2018.

[45] D. Gomez-Zara, M. Boon, and L. Birnbaum, “Who is the Hero, the Villain, and the Victim? Detection of Roles in News Articles using Natural Language Techniques”, in *23rd International Conference on Intelligent User Interfaces*, ser. IUI ’18, Association for Computing Machinery, Mar. 2018, pp. 311–315.

[46] C. Harris, “Searching for Diverse Perspectives in News Articles: Using an LSTM Network to Classify Sentiment”, in *IUI Workshops*, 2018.

[47] K. Kucher and A. Kerren, *Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights*. Apr. 2015, vol. 2015.

[48] R. Brath and E. Banissi, “Using Text in Visualizations for Micro/Macro Readings”, in *Proceedings of the ACM Intelligent User Interfaces Workshop on Visual Text Analytics*, Mar. 2015.

[49] M. Alharbi and R. S. Laramee, “SoS TextVis: An Extended Survey of Surveys on Text Visualization”, *Computers*, vol. 8, no. 1, 2019.

[50] S. Chen, L. Lin, and X. Yuan, “Social Media Visual Analytics”, *Computer Graphics Forum*, vol. 36, no. 3, pp. 563–587, Jun. 2017.

[51] A. Verhagen, *Constructions of Intersubjectivity: Discourse, Syntax, and Cognition*. Oxford University Press, Jun. 2007.

[52] G. Mishne and M. Rijke, “MoodViews: Tools for Blog Mood Analysis.”, in *Conference: Computational Approaches to Analyzing Weblogs*, Jan. 2006, pp. 153–154.

[53] H. Saif, M. Fernandez, and H. Alani, “Contextual Semantics for Sentiment Analysis of Twitter”, *Information Processing & Management*, vol. 52, Mar. 2015.

[54] S. Baccianella, A. Esuli, and F. Sebastiani, *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Jan. 2010, vol. 10.

[55] S. Kim, I. Woo, R. Maciejewski, D. Ebert, T. Ropp, and K. Thomas, *Evaluating the Effectiveness of Visualization Techniques for Schematic Diagrams in Maintenance Tasks*. Jan. 2010, p. 40, 33 pp.

[56] F. Danner, “Interactive Visual Correlation of Events in Social Media and News”, Bachelor’s thesis, University of Stuttgart, 2020.

[57] B. Adams, D. Phung, and S. Venkatesh, “Eventscapes: Visualizing Events over Time with Emotive Facets”, in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM ’11, Association for Computing Machinery, Nov. 2011, pp. 1477–1480.

[58] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, “Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry”, in *2010 IEEE Symposium on Visual Analytics Science and Technology*, Oct. 2010, pp. 115–122.

[59] K. Kucher, R. M. Martins, C. Paradis, and A. Kerren, “StanceVis Prime: Visual Analysis of Sentiment and Stance in Social Media Texts”, *Journal of Visualization*, vol. 23, no. 6, pp. 1015–1034, 2020.

[60] A. Abbasi and H.-c. Chen, *Categorization and Analysis of Text in Computer Mediated Communication Archives Using Visualization*. Jan. 2007, p. 18, 11 pp.

[61] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin, “TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 280–289, Jan. 2016.

[62] E. Guzman, P. Bhuvanagiri, and B. Bruegge, “FAVe: Visualizing User Feedback for Software Evolution”, in *2014 Second IEEE Working Conference on Software Visualization*, Sep. 2014, pp. 167–171.

[63] J. Bertin and W. J. Berg, *Semiology of Graphics: Diagrams, Networks, Maps*. ESRI Press : Distributed by Ingram Publisher Services, 2011.

[64] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, “Empirical Studies in Information Visualization: Seven Scenarios”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.

[65] S. Carpendale, “Evaluating Information Visualizations”, in *Information Visualization: Human-Centered Issues and Perspectives*, ser. Lecture Notes in Computer Science, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds., Springer, 2008, pp. 19–45.

[66] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko, “A Heuristic Approach to Value-Driven Evaluation of Visualizations”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 491–500, Jan. 2019.

[67] J. Stasko, “Value-Driven Evaluation of Visualizations”, in *Proceedings of the Fifth Workshop on beyond Time and Errors: Novel Evaluation Methods for Visualization*, ser. BELIV ’14, ACM, 2014, pp. 46–53.

[68] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. Keim, *State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams*. Jun. 2014.

[69] A. D. Boumaiza, “A Survey on Sentiment Analysis and Visualization”, *Qatar Foundation Annual Research Conference Proceedings*, vol. 2016, no. 1, 2016.

[70] B. Saket, A. Endert, and Ç. Demiralp, “Task-Based Effectiveness of Basic Visualizations”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2505–2512, Jul. 2019.

[71] L. Grammel, M. Tory, and M.-A. Storey, "How Information Visualization Novices Construct Visualizations", *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 943–952, Nov. 2010.

[72] K. Börner, A. Maltese, R. N. Balliet, and J. Heimlich, "Investigating Aspects of Data Visualization Literacy Using 20 Information Visualizations and 273 Science Museum Visitors", *Information Visualization*, vol. 15, no. 3, pp. 198–213, Jul. 2016.

[73] W3 Schools. (2021). "HTML Color Names", [Online]. Available: `https://www.w3schools.com/colors/colors_names.asp` (visited on Sep. 5, 2021).

[74] J.-D. Fekete and J. Freire, "Exploring Reproducibility in Visualization", *IEEE Computer Graphics and Applications*, vol. 40, no. 5, pp. 108–119, 2020.

[75] Linnaeus University. (2021). "Validity Definition by the Department of Computer Science and Media Technology", [Online]. Available: `https://coursepress.lnu.se/subject/thesis-projects/validity/` (visited on Sep. 5, 2021).

[76] M. L. Waskom, "Seaborn: Statistical Data Visualization", *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[77] L. Zhou and C. D. Hansen, "A Survey of Colormaps in Visualization", *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 8, pp. 2051–2069, Aug. 2016.

[78] M. Harrower and C. A. Brewer, "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps", *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, Jun. 2003.

[79] Seaborn User Guide and Tutorial. (2021). "Choosing Color Palettes: Qualitative Color Palettes", [Online]. Available: `https://seaborn.pydata.org/tutorial/color_palettes.html#qualitative-color-palettes` (visited on Sep. 5, 2021).

A Appendix 1

This section of the appendix lists all figures of the survey. The caption of each figure shows the corresponding question asked to the participants.

List of Appendix Figures

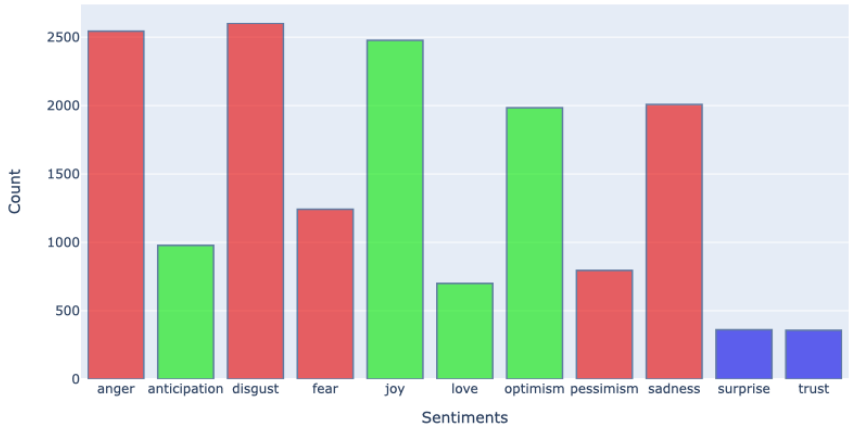


Figure A.1: Do you think that the emotions surprise and trust have the right color? How many positive sentiments can you identify in the above graph?

Figure A.2: The above graph shows a distribution of sentiment intensity scores. At what point does the data look abnormal?

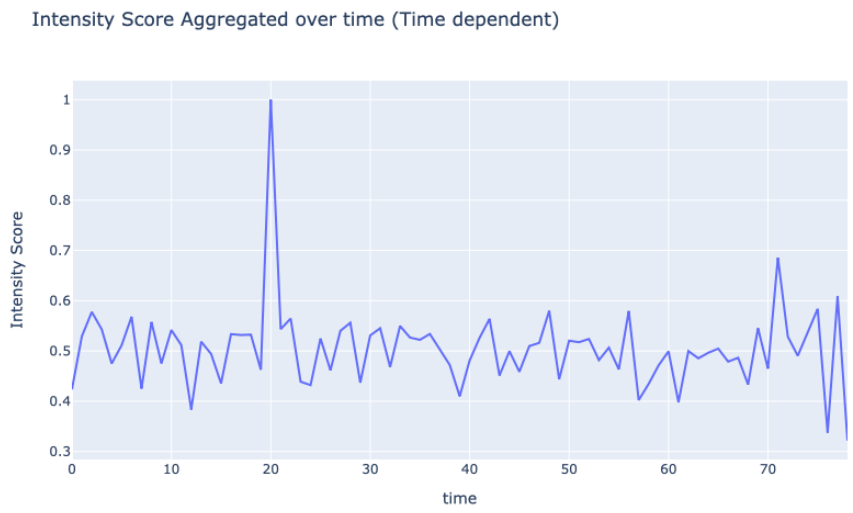


Figure A.3: At which timestep does the data look abnormal?

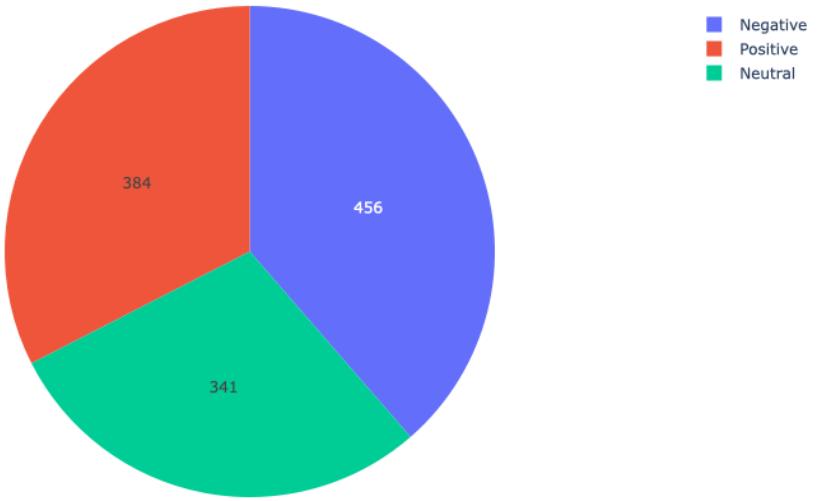


Figure A.4: What looks abnormal in the pie chart?

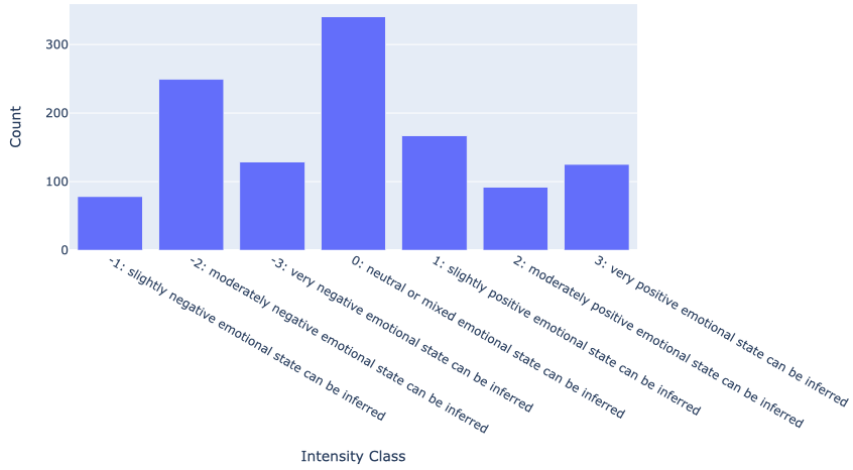


Figure A.5: How many different sentiments are shown in the graph?

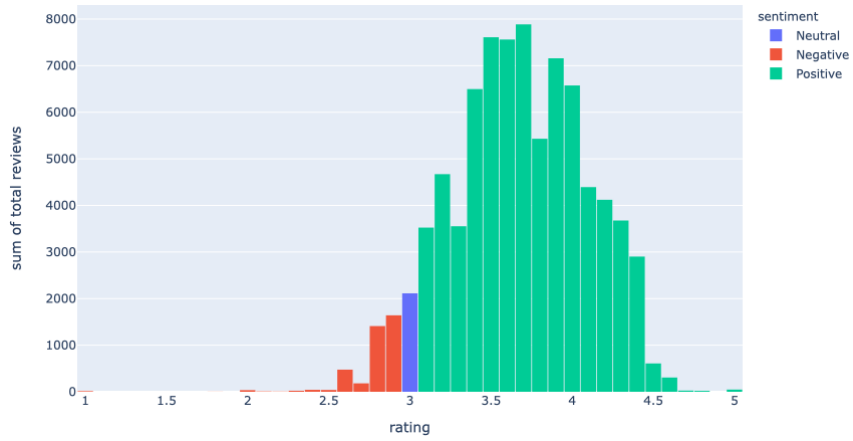


Figure A.6: What sentiment cluster has the largest sum of total reviews?

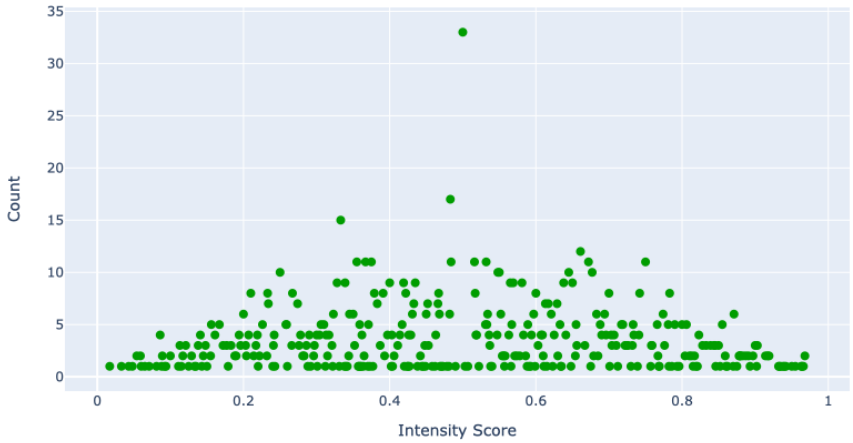


Figure A.7: Are you able to identify emotion clusters in the above graph?

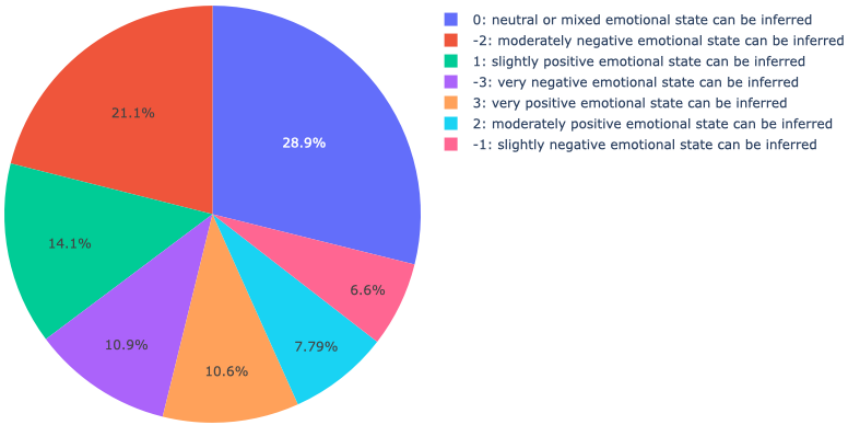


Figure A.8: What are the three main clusters of emotional states?

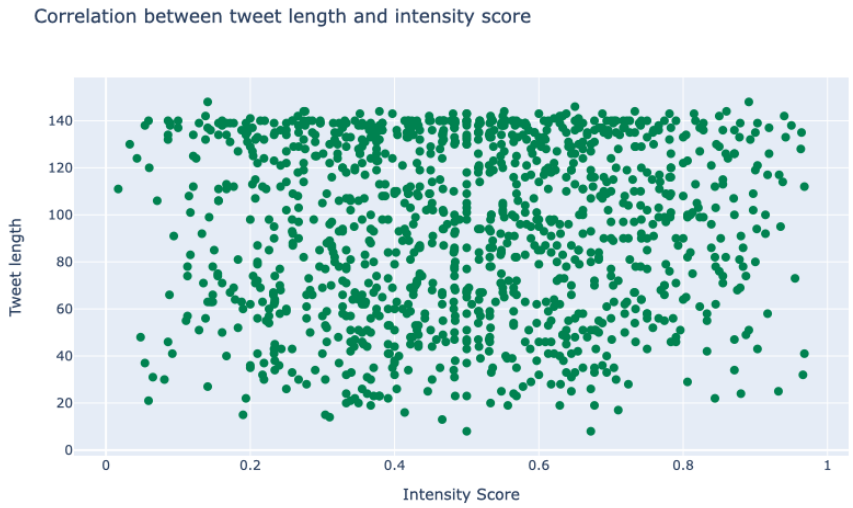


Figure A.9: Does Intensity Score have a correlation to the total tweet length?

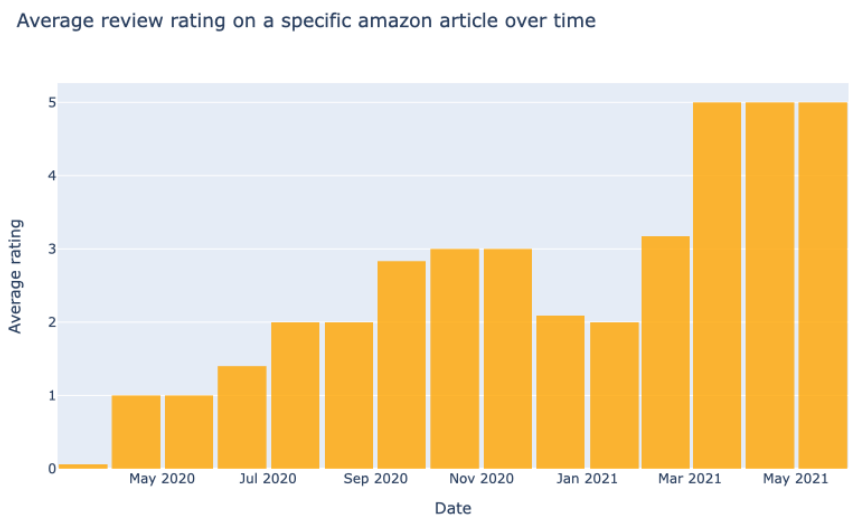


Figure A.10: Is there a correlation between time and average rating?

Figure A.11: From the above graph, can you tell that there is a (weak) correlation between ratings and the amount of reviews?

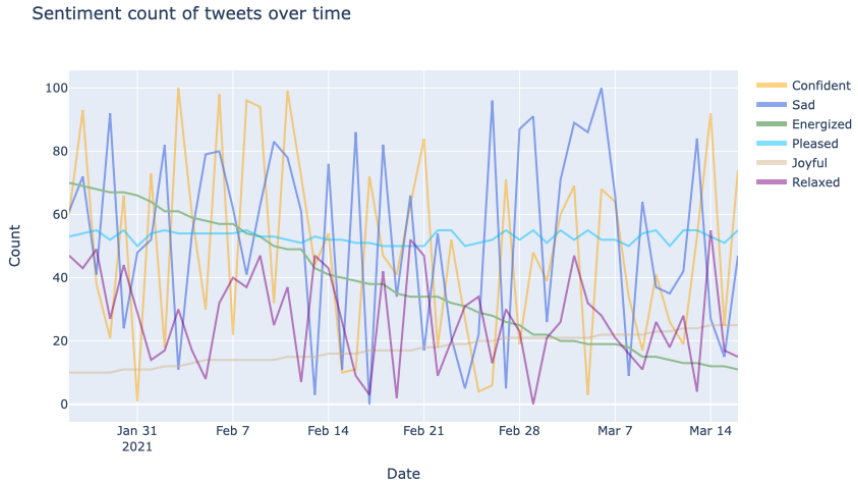


Figure A.12: What type of emotion is constantly decreasing in count over time?

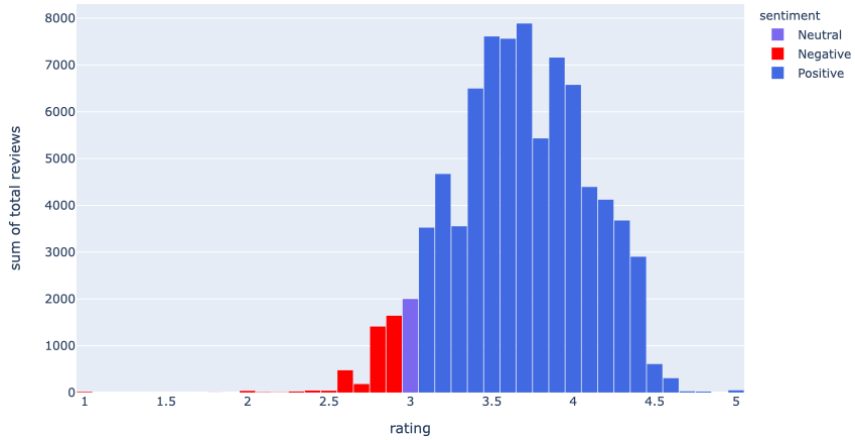


Figure A.13: How many neutral emotions can be identified in the graph?

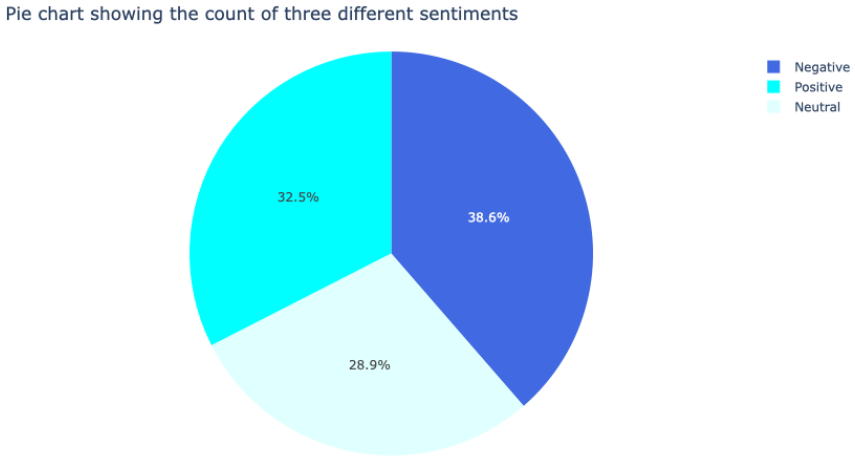


Figure A.14: What is the total percentage of positive and negative sentiments without neutral sentiments?

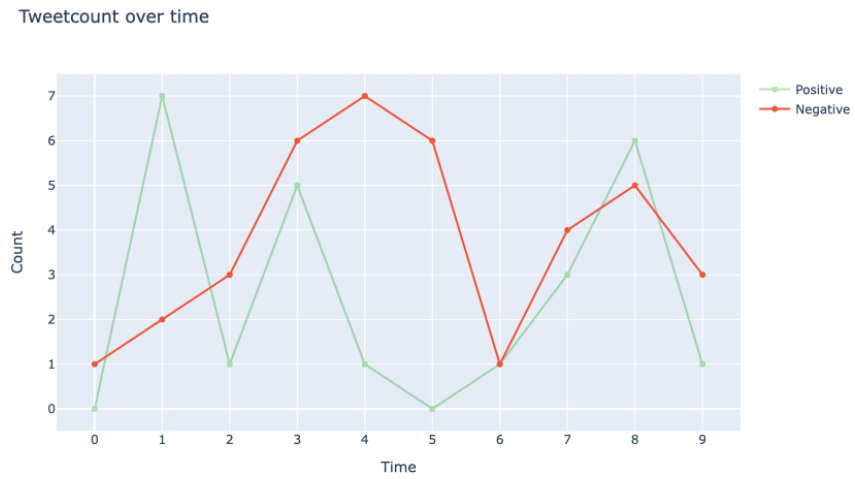


Figure A.15: How many negative tweets have been published since timestep 7?

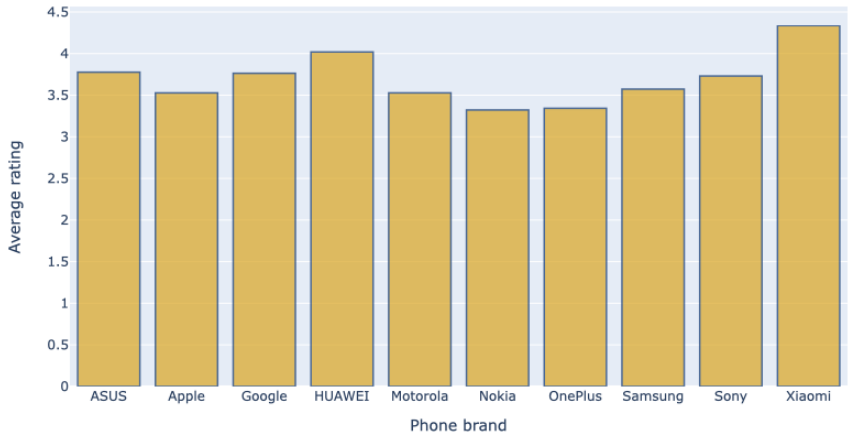


Figure A.16: What phone brand has the lowest average rating?

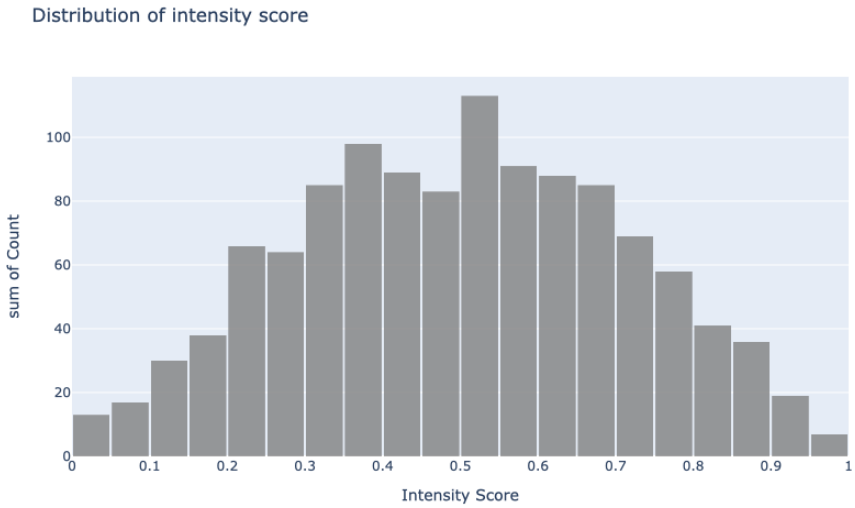


Figure A.17: What is the intensity score of the lowest sum of Count?

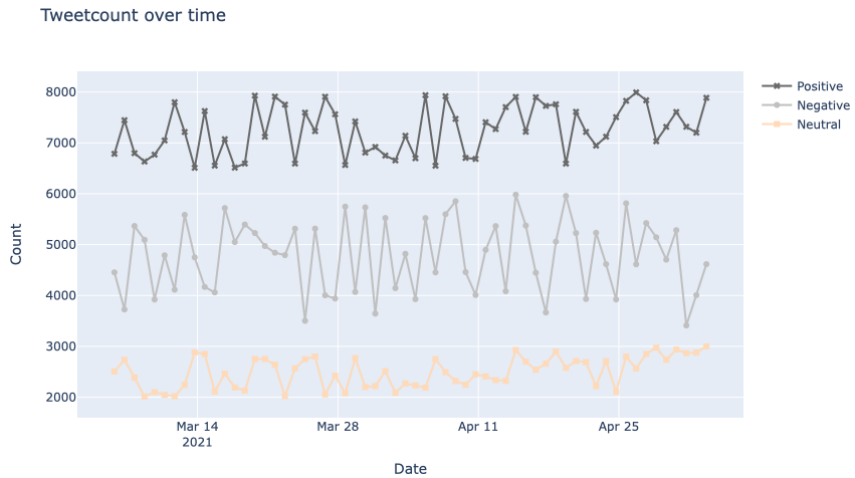


Figure A.18: At what time window was the day with the maximum number of negative tweets?

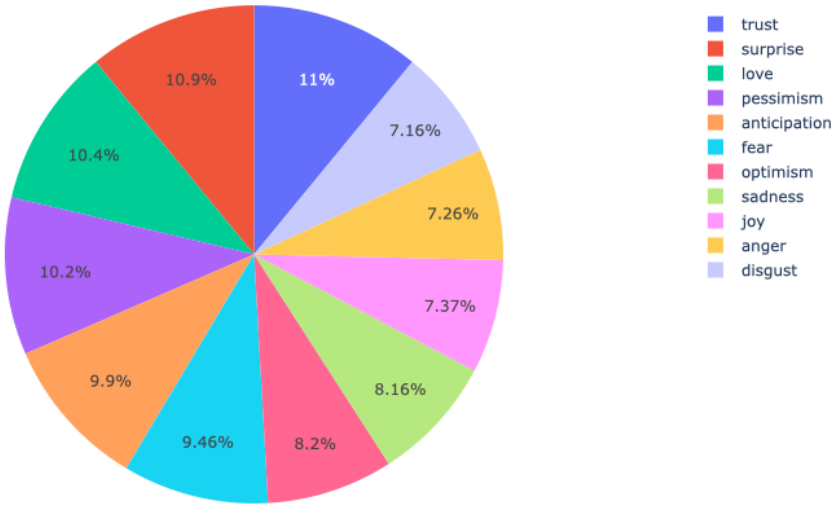


Figure A.19: Which two emotions have the most occurrences?

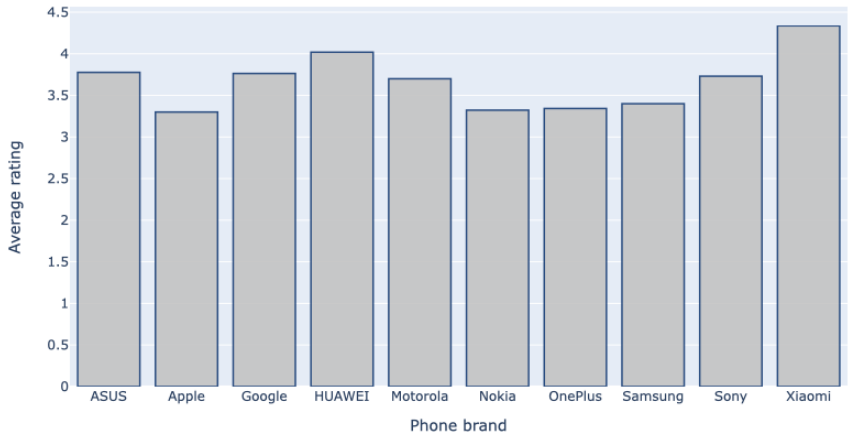


Figure A.20: How many phone brands have an average rating between 3 and 3.5?

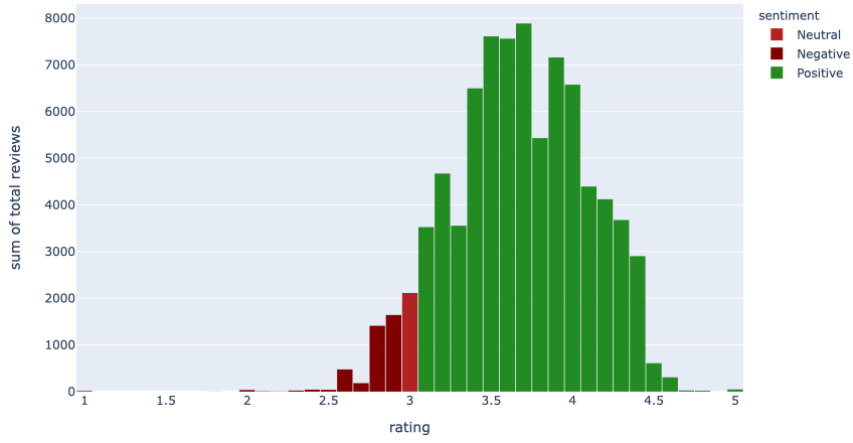


Figure A.21: What rating range has between 6000 and 8000 total reviews?

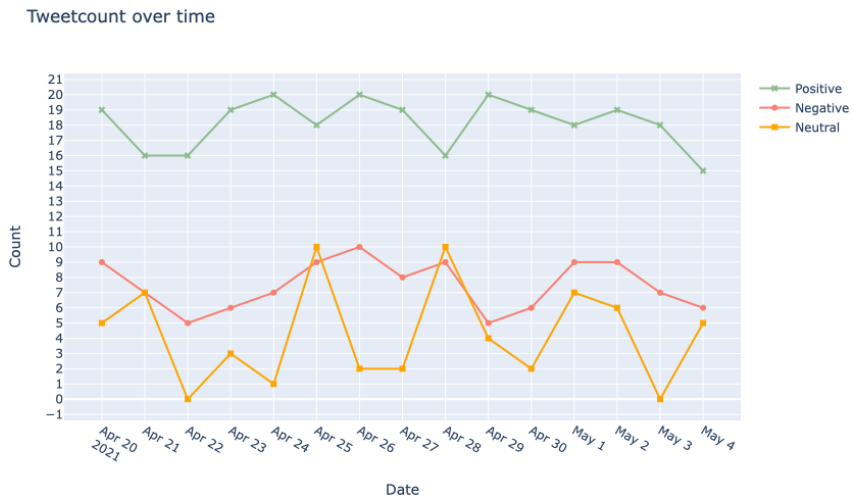


Figure A.22: How many neutral tweets were posted between Apr 26 and Apr 28?

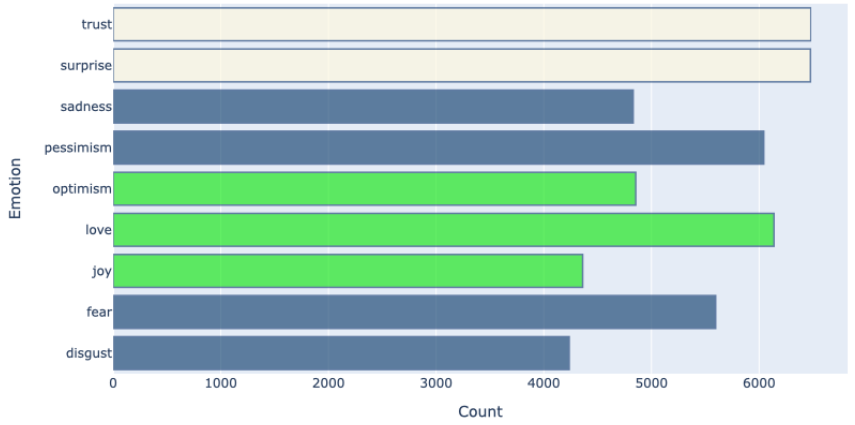


Figure A.23: How many different emotions are shown in the graph?

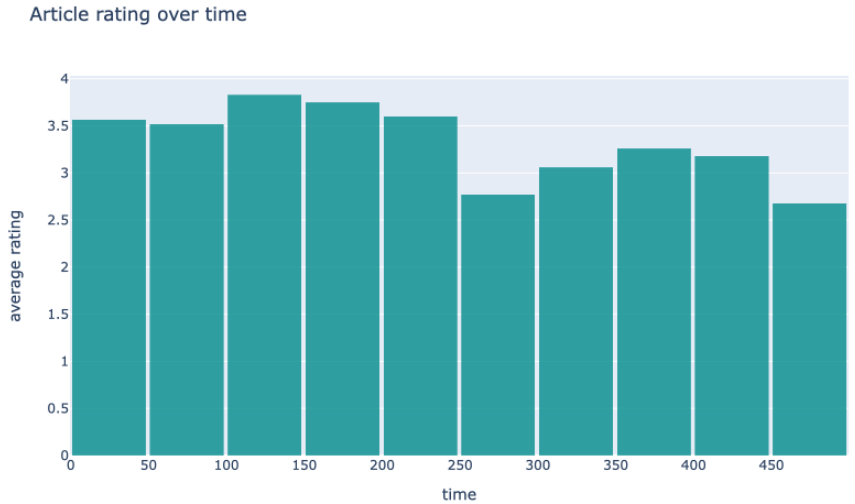


Figure A.24: What is the average rating of datapoints between timestep 250 and 300?

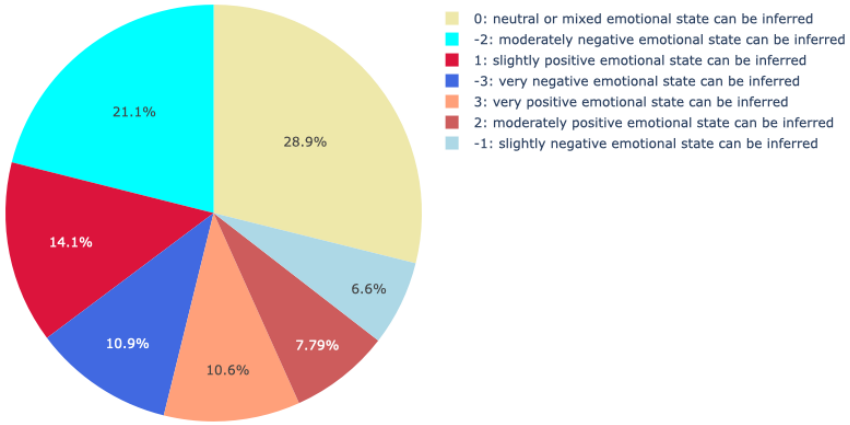


Figure A.25: What sentiment is the most present? How many emotions have a total count between 6% and 8%?

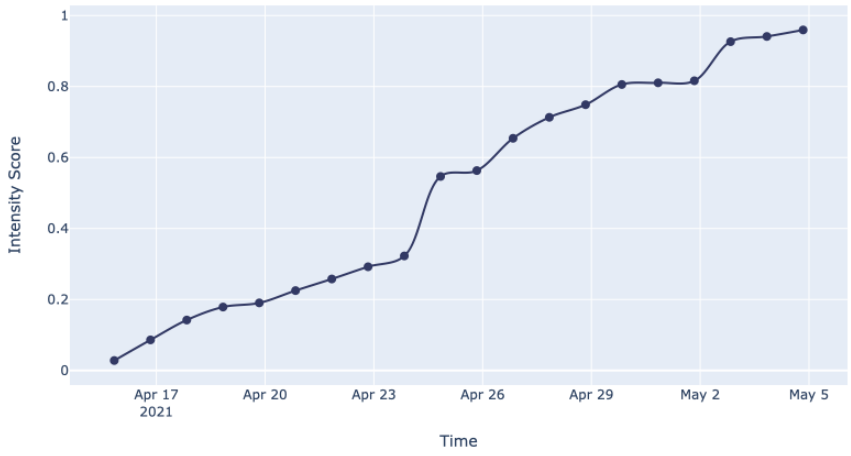


Figure A.26: What intensity score does the data point corresponding to May 2 have?

B Appendix 2

This part of the appendix lists the survey's raw results.

List of Appendix Tables

Table B.1: Survey questions with the correct answers and the amount of correct and wrong responses.

Question	Correct answer	Don't know	Correct	Wrong
What type of emotion is constantly decreasing in count over time?	Energized	0	50	0
What phone brand has the lowest average rating?	Nokia	0	50	0
What intensity score does the data point corresponding to May 2 have?	0.8	0	50	0
Which two emotions have the most occurrences?	trust and surprise	0	49	1
The above graph shows a distribution of sentiment intensity scores. At what point does the data look abnormal?	0.8-0.85	1	49	0
What is the total percentage of positive and negative sentiments without neutral sentiments?	71.1%	1	49	0
How many phone brands have an average rating between 3 and 3.5?	4	0	49	1
At which timestep does the data look abnormal?	20	1	48	1
How many emotions have a total count between 6% and 8%?	2	2	48	0
What cluster has the largest sum of total reviews?	Positive	0	48	2
How many different emotions are shown in the graph?	9	1	47	2
What is the average rating of datapoints between timestep 250-300?	2.5-3	3	47	0
What sentiment is the most present?	Neutral or mixed emotional state	2	44	4
How many neutral emotions can be identified in the graph?	2000	3	44	3
What is the intensity score of the lowest sum of Count?	0.95-1	3	43	4
Is there a correlation between time and average rating?	Yes, increasing average rating	6	40	4
How many different sentiments are shown in the graph?	7	2	40	8
At what time window was the day with the maximum number of negative tweets?	Apr 11-Apr 25	3	37	10
Are you able to identify emotion clusters in the above graph?	No	11	37	2
How many neutral tweets were posted between Apr 26 and Apr 28?	14	4	36	10
Does Intensity Score have a correlation to the total tweet length?	No correlation	14	36	0
How many negative tweets have been published since timestep 7?	12	7	34	9
What are the three main clusters of emotional states?	positive, negative and neutral	8	30	12
From the above graph, can you tell that there is a (weak) correlation between ratings and the amount of reviews?	No	12	12	26
How many positive sentiments can you identify in the above graph?	None	1	/	/
What looks abnormal in the pie chart?	None	2	/	/
What rating range has between 6000 and 8000 total reviews?	None	3	/	/
Do you think that the emotions surprise and trust have the right color?	None	1	/	/