COMPUTATIONAL APPROACHES FOR SEMANTICS-AWARE TYPOGRAPHICAL CHOICES

by

TUGBA KULAHCIOGLU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Gerard de Melo

and approved by

New Brunswick, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

Computational Approaches for Semantics-Aware Typographical Choices

By Tugba Kulahcioglu

Dissertation Director:

Gerard de Melo

Typographic signals carry strong semantic connotations, e.g., they may convey *excitement*, *anger* or even *sweetness*, which empowers them to affect almost any aspect of life, from perception of an email, to the perceived sweetness of a cup of coffee. This thesis explores some of the possibilities that can be offered by computational approaches to support users in understanding and taking advantage of this impact. More specifically, the focus is on learning font semantics from crowdsourced and Web data, and using this information to facilitate font search and recommendation. Among the novel contributions are the use of CNN-based embeddings to represent fonts in attribute learning, leveraging emotion theories and lexical relations to infer font semantics, a multimodal font search method that allows specifying a reference font together with the target semantic additions, enabled by a cross-modal representation of fonts and words, and the proposal of *affect-aware word clouds* that let users specify a target emotion, which is used to recommend fonts and color palettes with congruent affective connotations.

Acknowledgements

I gratefully acknowledge the support I received during the pursuit of this dissertation.

I would first like to thank my advisor, Professor Gerard de Melo, for his continuous support. I am particularly grateful for his invaluable technical advice, and for the freedom of working on the research problems of my own choice.

I am thankful for the support of all my friends. I want to specifically thank Min-Jeong Yang for useful discussions.

I also owe special thanks to our program coordinator, Michelle Walezak, for her friendliness and tremendous help.

Finally, I would like to thank my mother, Nalan Yeter, for her constant support and encouragement, and my sister, Nil Tugce Kulahcioglu, for inspiring me to work on this interesting topic, and for supporting my research with endless discussions and ideas.

Table of Contents

Chapter 1

Introduction

Typographical decisions, such as the choice of font and of stylistic features, determine the specific means through which a textual message is conveyed. They have the power to weaken, strengthen, or even alter the message being communicated. This chapter discusses our motivation to computationally analyze and apply font semantics, and provides an overview of the thesis together with main contributions.

1.1 Motivation

Paralinguistic signals are additions to the language units and structures that can modulate the meaning in many ways [1]. For example, a particular tone of voice could dramatically alter the underlying conceptual meaning of a sentence [2]. Typically associated with verbal language, paralinguistic phenomena also occur with written language [1; 3]. Emoji, font, color, and animation [4] are examples of textual paralinguistic signals [5]. The following two sentences convey different affective meanings using the same linguistic elements but different paralinguistic elements: “John is back in town :)”, “John is back in town :(” [1].

As a paralinguistic signal, typography received a significant interest from psychology and marketing research over the years [6; 7; 7–19]. The choice of font has been shown to affect the perception of the text as well as of the author or of the brand being advertised [13; 14]. It is widely agreed that different typefaces possess different *personas* in terms of their perceived associations and connotations [20–22]. These range from perceptions of *attractiveness* to evoked affective states such as *happiness* all the way to associations with *confidence* or even *laziness*. Understanding these latent semantic connections is a crucial precondition to using fonts adequately and effectively.

Having an important role in human communication, typography has not yet received adequate interest from the computer science community. To date, most software tools do not provide additional support for font selection. Among the tens of thousands of fonts available today, users are expected to go through each font to pick the one that serves their use case best. It is not uncommon for people to spend several minutes looking for the right font, but finally ending up using the default one [23]. Furthermore, as the number of available fonts keeps increasing, the task of font selection is becoming more challenging, especially for graphic designers, as their profession calls for such decisions to be made on a regular basis.

This thesis aims to contribute to the literature by developing methods to computationally analyze textual paralinguistic signals and by improving the user experience through the integration of these methods to the applications, with a special focus on fonts. Specifically of interest is the semantics of these signals rather than aesthetics [24] or readability [25], although sometimes these concepts are also explored as a part of the semantics (for example the attributes “clumsy” and "legible" in [21]).

1.2 Challenges and Contributions

This section discusses the challenges that are tackled in the scope of this thesis, and presents our contributions proposed to overcome these challenges.

1.2.1 Learning Font Semantics

The following discusses the challenges and the contributions associated with learning font semantics from existing font–tag connections.

Challenges. Existing methods use machine learning to learn models that can predict font semantics. Crowdsourcing [21], surveys [22], and Web data crawling [26] are currently used as primary methods to obtain a reference dataset that labels fonts with regard to semantic attributes such as *happy*, *thanksgiving*, or *pixel*. The challenges associated with learning from these different datasets vary. Crowdsourcing and surveys yield datasets that are clean and complete (a fairly accurate labeling of every font with regard to every attribute) but small (tens-to-hundreds of fonts, tens of attributes). In contrast, crawled Web data is large (thousands of fonts, thousands of attributes) but noisy (missing many font–attribute connections).

Contributions. Learning font semantics is an important focus of this thesis that is addressed in all technical chapters. In Chapters 3 and 4, we propose using a non-parametric learning algorithm, namely k-nearest neighbors (k–NN), to learn from small but high-quality crowdsourced data. In these chapters, for the first time in the literature, we use font embeddings extracted using pretrained CNN networks, and pretrained word embeddings to learn font semantics. In Chapter 5, we create, for the first time, a cross-modal representation of fonts and words, which helps overcome the challenges associated with large Web datasets that suffer from missing font–tag connections. As we discuss in the following section, the cross-modal representation also enables a set of novel interaction mechanisms with the fonts. In addition to using machine learning, in Chapters 4 and 6, we propose novel methods to infer font semantics by making use of emotion theories and lexical relations. As an example, in Chapter 4, we obtain values for the tag *optimism*, based on the font's existing connections with the tags *anticipation* and *joy* following Plutchik's Wheel of Emotion [27].

1.2.2 Font Search

The following outlines the challenges of searching for fonts using semantic tags, and discusses the contributions of the thesis on this topic.

Challenges. Users typically search fonts using semantic tags [28; 29]. A limited tag vocabulary and missing font–tag associations are primary challenges associated with this task. Even with a sufficiently large vocabulary and complete tags, users still need to spend significant time to find a font that meets their needs, and which is also *unique*, something necessary to better differentiate their design product from those of competitors. Users report that most of the time they find fonts that are very close to their ideal font, but "slightly off" in certain characteristics [30].

Contributions. In all chapters, our methods that are proposed to learn font semantics eventually target improving semantic font retrieval by learning more font–tag connections from existing ones. Beyond this general contribution, in Chapter 5, we propose a novel multimodal font discovery method that allows searching for fonts using a reference font and the targeted modifications to reach an ideal font. It aims to satisfy users' need to slightly modify a font to reach the ideal one. This is enabled by the cross-modal representation of fonts with words. The cross-modal representation enables other creative ways to discover fonts as well, such as using multiple semantic tags at once. The challenge to work with a limited vocabulary is also tackled by enabling search using any keyword(s) that are in the word embedding dictionary, which typically includes millions of words. All these improvements and supporting mechanisms together make it easier to deal with a vast number of fonts, increasing the chance to discover fonts that meet users' need and that are also *unique*.

1.2.3 Font Recommendation

The challenges associated with making successful typographical recommendations, and the contributions of this thesis to facilitate such choices are discussed below.

Challenges. Applications used for design, such as word processing and graphic design tools, require typographical choices but provide very limited built-in support. This leaves users alone with figuring out which font to choose, which can be a big challenge especially for non-expert users whom might not be aware of the impact of such choices, nor possess the knowledge required to make the right ones.

Contributions. We show that integrated recommendation mechanisms in end-user tools can facilitate the tasks that involve typographical choices. In Chapter 4, we recommend fonts for an English lexicon through crowdsourced emotion connections of words, and the connections of fonts to this small set of emotions. We illustrate how this font lexicon can be integrated into a poster-design tool to support font selection. As another application area, in Chapter 6, we develop a tool that matches congruent fonts and colors for word clouds based on the user-specified target emotions.

1.3 Publications

Following is a list of the publications that this thesis research is based on.

Kulahcioglu, Tugba, de Melo, Gerard. Predicting Semantic Signatures of Fonts. In Proceedings of the 12th IEEE International Conference on Semantic Computing (ICSC 2018), pp. 115–122.

Kulahcioglu, Tugba, de Melo, Gerard. FontLex: A typographical lexicon based on affective associations. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA).

Kulahcioglu, Tugba, de Melo, Gerard. Semantics-aware typographical choices via affective associations. Language Resources and Evaluation, (2020). Springer.

Kulahcioglu, Tugba, de Melo, Gerard. Paralinguistic Recommendations for Affective Word Clouds. In Proceedings of the 24th ACM International Conference on Intelligent User Interfaces (IUI 2019), pp. 132–143.

Kulahcioglu, Tugba, de Melo, Gerard. Affect-Aware Word Clouds. ACM Transactions on Interactive Intelligent Systems (2020).

Kulahcioglu, Tugba, de Melo, Gerard. Fonts Like This but Happier: A New Way to Discover Fonts. In Proceedings of the 28th ACM International Conference on Multimedia, (MM 2020).

1.4 Summary

Chapter 2 discusses related work, and then the following chapters proceed to present the aforementioned contributions of this thesis in detail. More specifically, in Chapter 3, we analyze a crowdsourced font semantics dataset and learn a method to predict font semantics for a much larger dataset. In Chapter 4, we use word embedding similarities and emotion-based connections to infer font semantics for an English lexicon, and demonstrate a potential use of the lexicon for a poster-design application. In Chapter 5, we obtain cross-modal representation for fonts and words that is used to propose a new multimodal font discovery method. In Chapter 6, we present an affect-aware font and color palette selection methodology for word clouds that aims to facilitate congruent typographical choices. Finally, Chapter 7 concludes the thesis.

Chapter 2

Background and Related Work

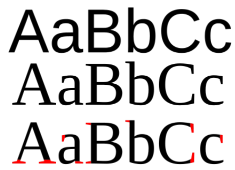
In this chapter, we provide background information on typography, and review related work on the impact of typographical signals and on font semantics. The related work on *color semantics* and *word clouds* are discussed in Chapter 6, as the discussed studies are not directly tied to font semantics, which is our main focus.

2.1 Typography

The written communication of humankind is thought to have started with *pictographs* that are basic images representing objects such as a *house*. *Ideographs* expanded upon this by associating images with ideas, or abstract concepts in general (e.g a *skull* representing *danger*). Then came *alphabets* that use symbols to represent the sounds of speech instead of ideas or objects [31].

Type refers to any specifically shaped reproducible such symbol [32], and consequently, *typography* is the art and technique of arranging type to optimally communicate the intended message. The arrangement of type involves typeface selection, and specification of its attributes such as point sizes, line lengths, line spacing, letter spacing, and kerning [33].

A *typeface* represents a specific design for a set of symbols from an alphabet [31]. Typefaces are usually classified into categories based on their shared visual characteristics. Two of the most common such categories are *serif* and *sans-serif*. *Serif* typefaces have a small stroke attached to the end of larger lines in a character. *Sans-serif* typefaces, in contrast, do not have such small lines, as demonstrated in Figure 2.1a. Another distinct category of typefaces are *handwriting* typefaces, also referred to as *script*, which use handwriting-like fluid lines. Figure 2.1b shows a *monospace* typeface, in which each



(a) Sans-serif (1st) and serif (2nd and 3rd) typefaces. Figure obtained from [34].

(b) Proportional and monospace typefaces compared. Figure obtained from [35].

(c) Sample dingbat fonts: "Journal Dingbats 2" (top row) and "Journal Dingbats 4" (bottom row).

Figure 2.1: Examples of font categories *sans-serif*, *serif*, *monospace* and *dingbat*.

symbol occupies the same amount of horizontal space. *Display* typefaces are designed to be decorative and attention-grabbing, hence result in a low level of legibility when used for body text. They are frequently used for headlines [32]. Finally, *dingbat* fonts are fonts that consist entirely of symbols instead of alphabetical or numerical characters (examples in Figure 2.1c). They are used for decorative or symbolic purposes.

An important point to note is the difference between a *typeface* and a *font*. While a *typeface* refers to a specific design, *font* refers to the actual file delivered to the user, possibly with certain stylistic effects, such as *italic* or *bold*. As these style differences can impact the semantics of a typeface, in this thesis we mostly focus on fonts rather than typefaces.

2.2 Impact of Typographical Signals

From *marketing* to *human-computer interaction*, an extensive set of studies show that fonts have a profound impact on customers' or users' perception. These studies can be clustered in two groups: *Stroop-style* studies and *survey-style* studies.

2.2.1 Stroop-Style Studies

Stoop-style studies ask users to respond to a given task both correctly and as quickly as possible, while measuring the response time. An example of such a study is the one by Lewis & Walker [6], in which users in one task are requested to press the left hand key if the words *slow* or *heavy* appear, and the right one if *fast* or *light* appears. They repeat such tasks with congruent (matching the underlying meaning or theme) and incongruent fonts, finding that congruent fonts significantly reduce the response times. Hazlett et al. [7] asked users to assess whether a displayed word is positive or negative. The results show that congruent font types result in faster responses to such tasks, similar to the *Stroop-Effect* caused by colors [36].

2.2.2 Survey-Style Studies

Survey-style experiments gather user ratings for semantic measures aiming to understand the impact of font characteristics on perception. These studies mostly focus on two areas: perception of text documents, and product packaging.

Text Documents. Juni & Gross [8] present two New York Times articles with two different fonts and solicit ratings from users. The results reveal that the same text is perceived as being funnier or angrier when read in a certain font compared to another. Shaikh [9] presents documents in three different fonts (congruent, incongruent, neutral), asking users to assess the personality of the document (e.g., exciting) and the personality of the author (e.g., trustworthy). The findings show strong effects for all three categories of fonts on the perceived personality of documents, whereas congruent and neutral fonts created similar perceptions of the authors' respective personalities. Hazlett et al. [7] displayed the same page with different fonts for 0.7 seconds each, asking users to describe the emotional tone of the page. They found that the latter is strongly influenced by the font type.

Shaikh et al. [10] investigate the effect of fonts on the perception of email and find that fonts with low congruency result in different perceptions of the email than when higher congruency fonts are invoked. A similar study on the perception of a company website [11] reveals that neutral/low congruency fonts negatively affect company's perception in terms of professionalism, believability, trust, and intent to act. Promotions advertised with right-slanted fonts have found to increase consumers' click through rates in email promotions, as well as retailer visit and purchase intentions [12].

Mobile Text Messaging. Choi et al. [37] investigate the impact of *angry* and *happy* typefaces on the perception of text messages, and compare this impact with the impact of using emoji or emotion words. Their findings reveal that, when used in congruency with the emotion of the text, text messages with happy fonts make the sender seem genuinely happy. An interesting comment from the users is that with a happy emoji at the end of the message, the receiver feels that the sender smiles after saying the words. In contrast, they feel that using the fonts the sender communicates the message with a happy voice all along. For the *angry* messages, some participants reported that they would not use a specific font or emoji if they were really in a very negative mood. All participants also agreed that negative words evoke stronger feelings than negative emoji or typeface. Finally, another user study using incongruent signals between text and font reveal that this situation is usually interpreted as humor or sarcasm.

Branding. Many studies in marketing analyze font effects, especially in packaging design. Fligner [13] shows that fonts associated with the attribute *natural* increase the perception of products with respect to *healthfulness* when used in their packaging, especially if the products' intrinsic (e.g., fat-free) and extrinsic cues (e.g., sold at Whole Foods Market) also support this. The experiments by Childers and Jass [14] show that semantic attributes of fonts affect user perception for both high and low engagement levels; and the effect of a font on the recall performance increases when other factors such as the picture used is consistent with the font. Through experiments using bottled water of a fictional brand, Van Rompay and Pruyn [15] as well found that the congruence between fonts and other design elements influence the perception of brand credibility, aesthetics, and perceived value. By means of three experimental studies, Giese and Parkman [38] find that brand personality perceptions, such as sincerity, sophistication and competence, are effected by the font used to display the brand name. The study also reveals that the impact of the font color on brand perception is independent of the impact of the font type itself.

Several studies investigated the cross–modal correspondence between taste and font shape. In [16], the participants are asked to rate their expectation on the sweetness and sourness of a product based on the presented packaging designs. The results support an association between sweetness and rounded fonts and sourness and angular fonts. Supporting these results, in [17], an association between sweet taste and rounded fonts are found, in addition to the association between bitter, salty, and sour tastes and angular fonts. A study on the perception of specialty coffee evaluated amateur consumers' expectation and experience on coffee acidity, sweetness, and their liking and purchase intent [18]. Their expectation was based on package labels with different fonts, and the experience after actually tasting the coffee. Fonts with angular characteristics were found to increase both the expectation and the experience of acidity and purchase intent. The study did not find an association between sweetness and round fonts as suggested by aforementioned studies. Another study conducted using different fonts on beer labels found a stronger association between sour and angular fonts, compared to rounded ones for both taste expectations and perceived taste [19]. A study on chocolate choices found that 75% of the customers choose the chocolate box with a congruent font rather than the box with an incongruent one [39]. The same study found similar effect on other product categories such as *car rental* and *soft drinks*.

Overall, these studies show that selecting congruent fonts has significant import on how the content, its authors, and associated entities such as products are perceived. Hence, it is crucial to develop techniques that aid in determining the semantic congruency of fonts.

2.3 Learning Font Semantics

There has been growing interest in computational approaches to analyzing fonts not just visually but with regard to their semantic associations. These approaches mainly make use of two data sources: crowdsourced/survey data and web data.

2.3.1 Crowdsourced and Survey Data

Through a crowdsourced study, O'Donovan et al. [21] associate 200 diverse fonts with 37 semantic attributes (e.g., *happy*). Specifically, they request participants to pick one of two presented fonts for a given attribute, and then aggregate these choices to assign fonts a series of scores for each attribute. The authors also learn semantic attributes for new fonts based on this crowdsourced data. An online survey [22] assesses the characteristics of 20 fonts with respect to 15 adjective-based scales (e.g., *stable–unstable*). Another survey study [40] learns font semantics using a CNN. Their training data consists of emotion values of 171 Japanese fonts for 12 emotion-representing adjective pairs that is collected based on the answers of the participants.

2.3.2 Web Data

Many font-focused websites [28; 29] allow contributors to tag fonts with attributes, some of which are more semantic than visual. Recent studies collect data from these resources to obtain a larger scale dataset. Chen et al. [26] use a Web dataset of 18.8K fonts and 2K tags and propose a generative feature learning algorithm together with an attention mechanism to infer font–tag connections used for font retrieval. The authors use rendered images of each character individually to train the network, as opposed to using a text with multiple characters, claiming that the global structure of a text image does not provide extra information for font retrieval purposes. Choi et al. work on Web data consisting of 9.3K fonts and 3.7K tags [41]. They allow font querying with any input word by mapping the input words to the predefined tag dictionary with the help of word-embedding based similarities.

2.4 Interaction with Fonts

In this section we provide the methods that are currently available to users to interact with the fonts, or proposed by academic research as useful potential tools. We exclude the extensively researched image-processing topic of font recognition [42–46] and focus on font search, recommendation and design methods as they are related to the semantics of the fonts.

2.4.1 Font Search

We survey the main methods to discover fonts: name/category based search, keyword based search, and similarity based search.

Search by Name/Category. Most of the available word processing and graphic design tools allow users to search the fonts only by their names (e.g., Times New Roman), or via a predefined narrow set of categories (e.g., monospace) with a few semantic ones (e.g., fun). Users, thus, frequently consult external resources, such as the web-sites discussed below, to find fonts that meet their needs.

Attribute-based Search. Many web-sites provide tag-based font search powered by the font–tag associations provided by their users. Most of the time the designer of a font also enters tags when they upload a new font to Web. Although useful, these resources are limited by the tag vocabulary that is being used, and not every font is necessarily tagged with a sufficient number of tags. To overcome these challenges, several academic studies work on the problem of attribute-based retrieval of fonts [21; 26; 41; 47]. Most of these studies work with a limited tag vocabulary but extend the tag associations to new fonts [21; 26], while, through the use of word embeddings, in [41] the search space is extended to further words as well. In [48], the authors develop an inspiration tool that provides unexpected but useful font images or concept words in response to a user query.

Similarity Search. O'Donovan et al. [21] proposed a similarity-based font search in which users look for a font that is similar to a reference font. It is demonstrated in their study that human perception of similarity is closer to the semantic similarities of the fonts than the visual similarities, although both aspects contribute to the similarity perception. Wang et al. [44] propose a deep convolutional neural network (CNN) approach to help users identify the fonts employed in a photographic image, which is also proposed as a method to obtain font embeddings that can help assess font similarities. Lim et al. [47] present a font matching approach that retrieves fonts similar to a reference font for which users can decide the weights for three different similarity metrics: usage, personality, and shape. Qiu et al. [49] propose a method to map Latin and Japanese fonts that share similar semantic profiles.

Sketch-based Search. Ishibashi et al. [50] propose a sketch-based font retrieval method, which retrieves fonts similar to the sketch specified by the user. To avoid the need to draw the sketches from scratch, their method allows editing an existing font to search for fonts that are closer to the user's target font.

2.4.2 Font Recommendation

In this section we provide studies that seek to recommend fonts based on a specified text, image, or a font to be paired.

Text-Based Font Recommendation. Kawaguchi and Suzuki presents a system that recommends fonts, as well as colors, for book covers based on the content of the book [40]. The recommendation is enabled through an emotion vector that is obtained for both the fonts and the book text. Similarly, Chou and Lin [51] recommend Chinese fonts for emotion-bearing sentences based on the determined emotion of the text. In a recent study, a series of deep neural network models are explored to assess a short input text and to perform multi-label classification that select the best-fitting ones among 10 different display fonts [52].

Image-Based Font Recommendation. FontMatcher [53] recommends fonts for image captions, using a color–font mapping in the soft–strong and warm–cool dimensions, using a crowdsourced dataset [21]. Evaluated by a user study. the author found that their system outperform novice users. Similarly, in [41], a query-by-image approach is proposed for font selection. Zhao et al. [54] predict font typeface, color and size for web-design, by learning such connections using web-page screenshots and associated font metadata. Their neural network architecture also makes use of font embeddings generated using an autoencoder.

Emotion-Based Font Recommendation. Choi et al. [37] develop a messenger application prototype that lets the users specify an emotion before sending the message, giving them the option to use a font congruent with the selected emotion, which can be *happiness* or *anger*.

Font Pairing. In graphic design products, fonts are often used in pairs. These pairs have specific properties, e.g., share an overarching theme but have a pleasing contrast. This knowledge is rooted in graphic design principles and not easily available to novice users. An interesting font interaction method proposed is to search for a font pair given a specific reference font. FontJoy [55] is an online tool that uses deep convolutional font embeddings to achieve font pairing. Jiang et al. [56] present a font-pairing method based on font-pair data crawled from PDF documents on the Web.

Font Design. Recently, Wang et al. [57] proposed a computational font design method where the fonts are computationally designed and generated based on the user-specified attributes. Using both typographic and semantic attributes [21], their model is trained to achieve style transfer between two fonts conditioned on their corresponding attribute values. Other studies partially automate the design process by transferring the user's font style from a few input glyphs to a complete typeface [58; 59].

Chapter 3

Predicting Semantic Signatures of Fonts

In this chapter, we aim to induce *semantic signatures* that characterize the semantic attributes of large numbers of fonts.

One way to assess a wider range of fonts is to draw more general conclusions by unearthing connections and trends that apply broadly at a higher level, as in the study that found an association between *sweet* taste and *round* typefaces [16]. To this end, we first analyze the relationships of font categories and font styles with semantic attributes to determine to what extent simple correlations may exist. We find that certain attributes are indeed manifested most predominantly in specific categories and styles of fonts. However, this does not hold in general, as the interactions of visual features and semantic attributes are not always straightforward. and thus, this generic approach is not always feasible.

As a second step, overcoming the aforementioned challenges, we develop a semi-automatic computational approach to predict semantic signatures of fonts based on a small set of seed data. With this method, we extend crowdsourced data for 200 fonts to a large dataset, currently covering 1,883 fonts, which we make available online. Figure 3.1 presents the attributes used in this study, visualized using the fonts found to be most congruent by the extended dataset, excluding the ones from the crowdsourced seed data.[1]

We then proceed to analyze the resulting semantic resource, with the aim of assessing its quality as well as to derive insights regarding the potential of fonts to exhibit desired semantic attributes. The interactive visualization used to carry out this analysis is also made available online.

[1]The attribute *attention-grabbing* is shortened as *attention*.

Figure 3.1: Attributes used in this study, visualized using the font that is predicted to have the highest congruency[2].

The chapter is organized as follows. In Section 3.1, we analyze the semantic signatures of fonts based on font categories and styles. In Section 3.2, we present the method we use to induce signatures for an extended set of fonts, and evaluate its performance. Following this, in Section 3.3, we analyze the resulting dataset. We provide a discussion in Section 6.7 and then conclude in Section 6.8.

3.1 Semantic Signatures of Font Categories and Styles

Our first goal is to expose general associations between font categories or styles [32] and semantic attributes. As a starting point, we consider the crowdsourced data by O'Donovan et al. [21], which associates 200 fonts with 37 semantic attributes (e.g., *happy*, *formal*). We normalize their ratings to the [0,1] range, and derive attributes at the level of font categories, as well as for italic emphasis and font weights, to analyze their relationship with different attributes[3]. The results of this process are depicted in Figures 3.2, 3.3, 3.4, and 3.5. These plots are taken from an interactive visualization that we have established for the analysis of this data and made available online[4].

[2]Aiming to showcase different fonts, we used the next most congruent font if the most congruent font is already used for another attribute.

[3]We exclude 6 typographic attributes and use the remaining 31 attributes.

[4]Supplementary material can be accessed via http://gerard.demelo.org/fonts/

Category	Count	Average Avg.	Average SD	Max Avg.	Max Att.	Min Avg.	Min Att.
display	45	0.56	0.18	0.80	fresh	0.37	delicate
handwriting	18	0.63	0.15	0.84	gentle	0.29	boring
monospace	8	0.48	0.16	0.81	gentle	0.29	modern
sans-serif	85	0.49	0.14	0.87	gentle	0.24	clumsy
serif	44	0.53	0.16	0.89	gentle	0.19	bad

Table 3.1: Summary of statistics for the font categories. (SD: Standard Deviation, Att: Attribute, Avg: Average)

3.1.1 Font Categories

We begin by analyzing the relationship of five coarse-grained font categories.[5] Table 3.1 provides a high-level summary of these categories. Although both the averages and maximum scores appear to be close, and the scores of the respective attributes with the maximum average as well, a cursory glance at Figures 3.2 to 3.5 reveals that the distributions diverge significantly between particular font categories.

Display. Following the highlighted (yellow) lines in Figure 3.2, we observe that the *display* category appears to have the most scattered attribute scores. Across nearly all considered attributes, we find that its scores lie in a high range. This is also reflected in the summary table with a relatively high standard deviation value of attribute averages.

Handwriting. Figure 3.2 reveals that handwriting fonts appear to show a trend rather different from those of other font categories, especially for the attributes *artistic*, *charming*, *complex*, *dramatic*, *modern*, and *playful*, they score higher than others. The category also has strong associations for the attributes *fresh*, *friendly*, and *gentle*, which accords with the general trend. For *boring* and *strong*, in contrast, it has particularly low scores.

[5] These categories reflect historical origins and typographic properties. *Handwriting* typefaces are designed to create the impression of being hand-rendered. The characters of *monospace* typefaces occupy equal horizontal space. *Serif* typefaces have small lines attached to the end of the strokes in its characters, whereas *sans-serif* denotes typefaces lacking those attached lines. *Display* typefaces do not share typical typographic qualities other than a low level of legibility when used for body text, so they are reserved mostly for headings and other kinds of display purposes.

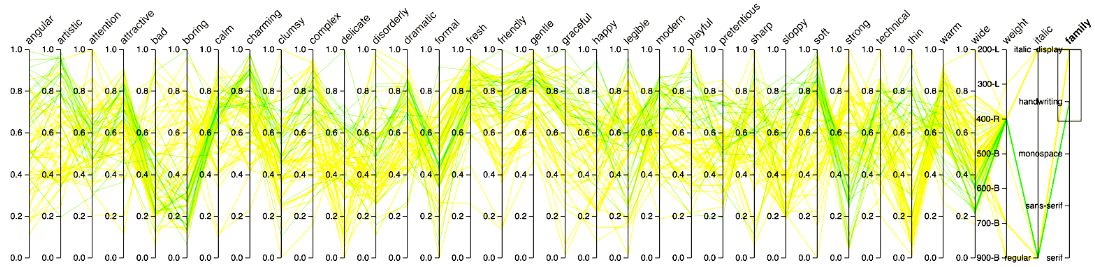


Figure 3.2: Semantic signature of the font categories *display* and *handwriting*. This shows the font association values for semantic attributes, and font categories and styles. Each line is colored based on the font categories (*display*: yellow, *handwriting*: green).

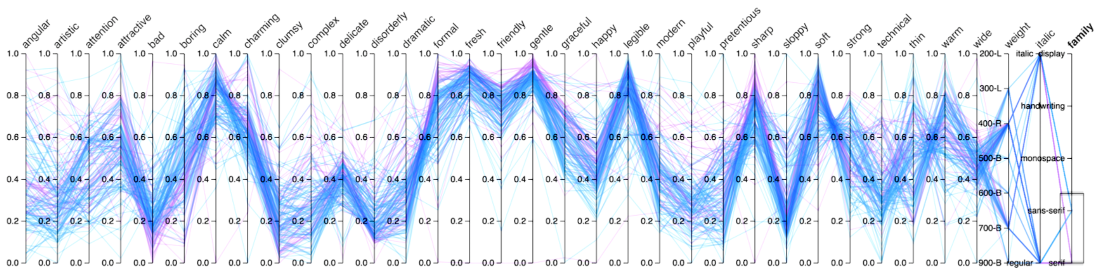


Figure 3.3: Semantic signature of the font categories *serif* and *sans-serif*. This shows the font values for the semantic attributes, and font categories and styles. Each line is colored based on the font categories (*sans-serif*: blue, *serif*: pink).

Monospace. Monospace fonts make up a very small number of instances in the data, only 8 out of 200. Their curve follows the general trend for *fresh* and *gentle* and *calm*, while having atypically high values for *boring*.

Sans-serif. This is the largest category in the dataset, with 85 members. The distribution of attributes can be analyzed in detail in Figure 3.3. We observe high association scores for *calm*, *formal*, *fresh*, *friendly*, *gentle*, *legible*, and *soft*, whereas for *bad*, *clumsy*, *disorderly*, *playful*, and *artistic*, we encounter lower values. The strongest association is for the attribute *gentle*, while the lowest score is seen for *clumsy*.

Serif. The serif font category has 44 samples in the dataset, and follows a similar pattern as *sans-serif*, except for showing slightly higher associations for *formal*, *gentle*, *friendly*, *happy*, and *sharp*, and slightly lower values for *bad*. Its highest score is for the attribute *gentle*, just as for sans-serif, while its lowest is for the attribute *bad*.

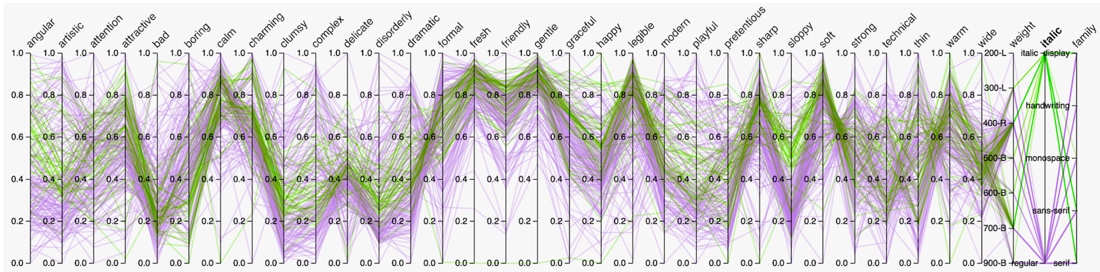


Figure 3.4: Semantic signature of the font style *italic*. This shows the font values for the semantic attributes, and font categories and styles. Each line is colored based on the italic emphasis of the fonts (*italic*: green, *regular*: pink).

		Average		Max		Min	
Style	Count	Avg.	SD	Avg.	Att.	Avg.	Att.
italic	42	0.56	0.16	0.87	gentle	0.26	bad
regular	158	0.52	0.18	0.84	gentle	0.32	bad
bold	59	0.52	0.15	0.86	fresh	0.27	disorderly
normal	127	0.54	0.18	0.85	gentle	0.31	bad
light	14	0.44	0.11	0.91	soft	0.14	pretentious

Table 3.2: Summary of statistics for the font styles. (SD: Standard Deviation, Att.: Attribute, Avg.: Average)

3.1.2 Font Styles

Table 3.2 summarizes statistics for the font style properties that we analyze: *italic* emphasis and *weight*. With the exception of the *light* font weight style, the values are very similar across all styles.

Figure 3.4 plots the distributions for fonts with *regular* (158 samples) and *italic* (42 samples) styles. For the attributes *artistic*, *complex*, *disorderly*, *dramatic*, and *playful*, *italic* seems to have mid-range values, whereas the *regular* style constitutes the high and low peaks. They both seem to peak for those attributes that also exhibit a general trend of having high values for the fonts in our data, such as *calm*, *fresh*, *gentle*, and *legible*. We found that among the font categories, *serif* fonts suffer the greatest impact when this style property is applied. The most *charming*, *attractive*, and *happy serif* fonts, for example, all use *italic* forms.

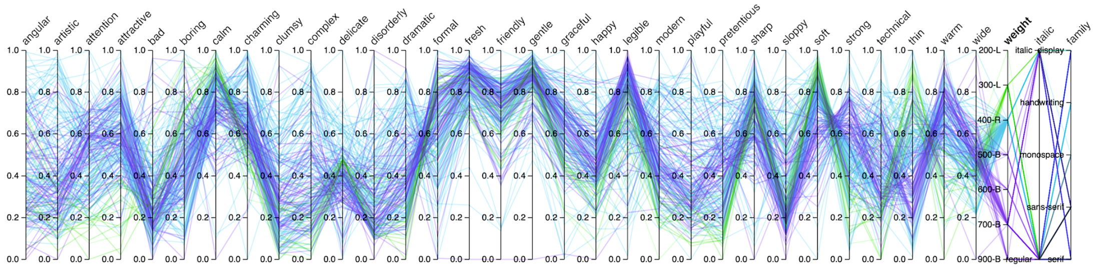


Figure 3.5: Semantic signature of the font style *weight*. This shows the font values for the semantic attributes, and font categories and styles. Each line is colored based on the weights of the fonts (*light*: green, *normal*: blue, *bold*: purple).

Figure 3.5 plots the distributions for fonts with different weights. Weights below 400 are considered *light* (14 samples), whereas the ones above 400 are considered *bold* (59 samples). The *normal* weight is assumed as 400 and consists of 127 samples for our data. The attributes *thin*, *soft*, and *calm* appear to have high values for fonts in a *light* style. Similarly, the attribute *warm* correlates with the *bold* style. They all seem to have peaks for the attributes *legible* and *gentle*. The least *happy* fonts are those that are *light*. Further analysis reveals that *sans-serif* shows strong interactions with *weight*, e.g., the *calmest* and *softest sans-serif* fonts use *light* forms, whereas *warm* and *legible* fonts use *bold* forms.

3.2 Large-Scale Semantic Signature Induction

We now proceed to produce a much larger-scale database of semantic signatures.

3.2.1 Method

We assume as input a set of fonts $\mathcal{F}$ described in terms of a set of font attributes $\mathcal{A}$. For this, we again rely on the previously used crowdsourced data by O'Donovan et al. [21], which describes a small set of 200 fonts. For a given font $f \in \mathcal{F}$, it provides scores in $[0, 100]$ for each attribute $a \in \mathcal{A}$. From this data, we derive $|\mathcal{A}|$-dimensional vectors $\vec{f} \in [0,1]^{|\mathcal{A}|}$ for each $f \in \mathcal{F}$, by transforming the dataset to consider the attributes for a given font while normalizing scores to $[0, 1]$.

Our aim is to predict $\vec{f'}$ for fonts $f' \notin \mathcal{F}$. To achieve this, we use k-nearest neighbors (k–NN) regression. The distance between two fonts, denoted as $d(f_i, f_j)$ is calculated using one of the similarity metrics described in the following subsection.

The unweighted k–NN approach uses the following formula, where $\vec{f_1}$ to $\vec{f_k}$ are attribute vectors for the closest k fonts in $\mathcal{F}$ according to a similarity metric.

$$\vec{f} = \frac{1}{k}\sum_{i=1}^{k}\vec{f_i} \tag{3.1}$$

The weighted k–NN approach generates weights using the following equation.

$$w_i = \frac{1}{k-1}\frac{\sum_{\substack{j=1\\ i\neq j}}^{k} d(f', f_j)}{\sum_{j=1}^{k} d(f', f_j)} \tag{3.2}$$

Subsequently, the weighted values are generated as follows:

$$\vec{f} = \sum_{i=1}^{k} w_i\vec{f_i} \tag{3.3}$$

3.2.2 Similarity Measures

To compute nearest neighbors, we consider four similarity metrics as alternatives.

The first option is to use typographic properties, obtained by parsing a font's glyph outlines to extract italics, thickness, size, area, orientation, stroke width, and spacing. We rely on existing data for this [21]. For some of these features, the data provides an average for all the characters, whereas for others, only selected characters are used.

The second option is to use a deep Convolutional Neural Network (CNN) to induce a font embedding space that captures font similarity [60]. For this, we rely on a model[6] that creates images by rendering a set of selected letters (L,a,s,e,g,d,h,u,m,H,l,o,i,v) in a grid, and then feeds them through a pretrained deep convolutional network. Finally, PCA is used for dimensionality reduction to obtain vectors that are compared in terms of cosine similarity. The dataset contains embeddings for 1,883 fonts.

[6] https://github.com/Jack000/fontjoy

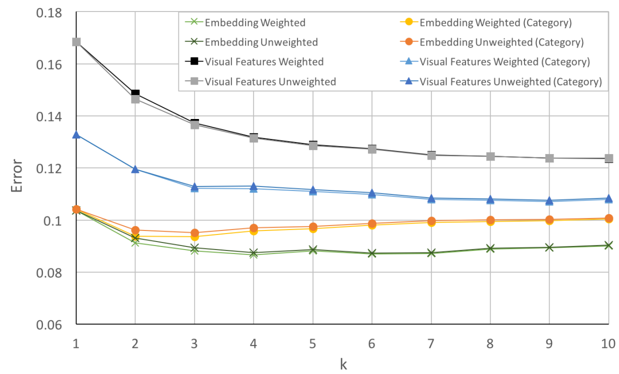


Figure 3.6: Error plots of four method–similarity measure combinations using different k values.

We consider two further alternative similarity measures that effectively restrict the candidate spaces of the above two measures to fonts having the same category as the input font. For example, for an input *handwriting* font, these measures regard all non-*handwriting* fonts as having a similarity of zero.

3.2.3 Evaluation

To evaluate this, for each f in $\mathcal{F}$[7], we predict $\vec{f}$ using $\mathcal{F} \setminus f$. We replicate the tests four times, for combinations of similarity measures and methods (weighted, unweighted). Comparing predicted $\vec{f}$ with the ground-truth $\vec{f'}$, an $|\mathcal{A}|$-dimensional vector $\vec{e}$ is calculated as:

$$\vec{e} = \vec{f'} - \vec{f}. \tag{3.4}$$

For each attribute, we then generate an error value e by averaging the absolute values of errors in $\vec{e}$. The test results are summarized in Figure 3.6. The error scores reported here are averages over e values across all $a \in \mathcal{A}$. The CNN embedding similarity metric results in a lower e for both the weighted and unweighted methods. *Category* based similarities led to slightly improved results for visual features, whereas they did not

[7]We use the 161 fonts that are common to all datasets.

show any improvements for the embeddings. The lowest error is obtained when $k = 4$ for the weighted version.[8]

attribute	e	**attribute**	e	**attribute**	e
fresh	0.051	strong	0.086	boring	0.097
gentle	0.051	attention	0.086	playful	0.098
delicate	0.057	bad	0.087	formal	0.099
wide	0.059	modern	0.089	warm	0.101
charming	0.062	legible	0.091	thin	0.101
friendly	0.063	disorderly	0.094	sharp	0.106
calm	0.072	attractive	0.095	angular	0.107
soft	0.076	dramatic	0.095	complex	0.108
graceful	0.077	pretentious	0.096	technical	0.111
sloppy	0.081	artistic	0.097		
happy	0.082	clumsy	0.097		

Table 3.3: Error averages for each attribute.

Table 3.3 lists the e value for each $a \in \mathcal{A}$ using the weighted embedding method where $k = 4$. The most successful predictions are made for *fresh*, whereas the least successful ones are for *technical*. The error scores lie in the narrow range between 0.05 and 0.11, whereas the full value range is between [0,1]. Analyzing these attribute-based error values together with the interactive visualization introduced in the previous section reveals that attributes with lower ranges have a lower degree of error, whereas attributes with high ranges tend to have greater levels. Another factor that appears to have an impact is the distribution of attribute values among different font categories. High-ranged values for which different sub-ranges are dense in certain categories seem to be associated with a lower error than ones with mixed such distributions.

Figure 3.7 shows how e changes with respect to $d(f', f)$. The likelihood of an error increases with increasing distance. However, there are also many cases in which the error is low despite high distances. It is also clear that certain font categories are easier to predict (such as *serif* and *sans-serif*) than others (such as *display*). Figures 3.8a and 3.8b show font samples with the highest and lowest e, respectively.

[8] We attempted to compare the error distribution of the typographical features against the CNN approach to explore to what extent these two metrics might provide complementary signals. However, they both seem to share a similar error pattern. Hence, it was not possible to obtain a significant improvement through a hybrid use of these metrics.

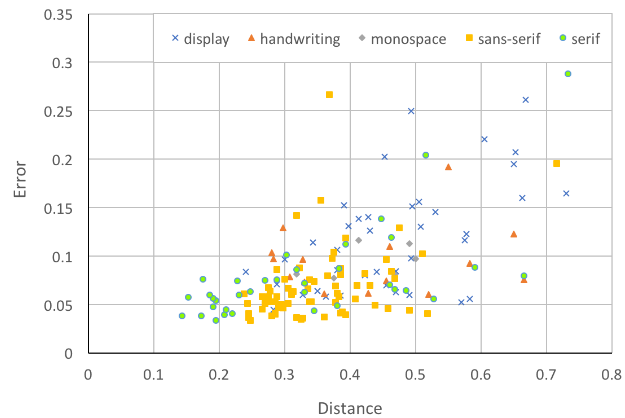
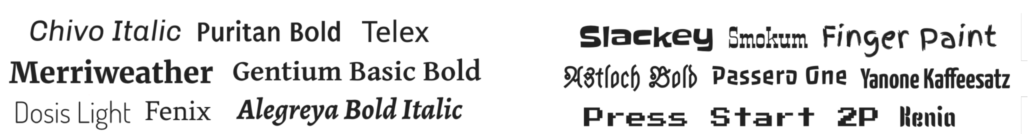


Figure 3.7: Scatter plot of the results relating distances of the samples and corresponding error values.



(a) Fonts with the lowest error averages.

(b) Fonts with the highest error averages.

Figure 3.8: Examples of fonts with the lowest and highest error averages.

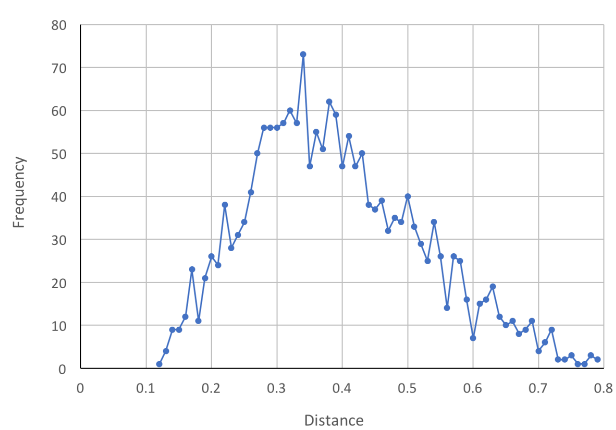


Figure 3.9: Distance distribution of the generated dataset.

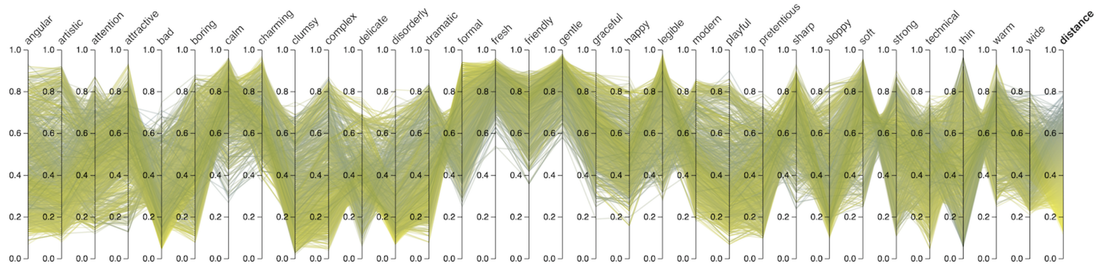


Figure 3.10: The semantic signatures of the fonts as generated by our method. This shows the distribution of the font values for the semantic attributes used. Each line is colored based on the distance value of the corresponding font.

3.2.4 Attribute Prediction

Finally, we predict $\vec{f}$ for all fonts covered by the CNN embeddings, using the weighted method with $k = 4$. Figure 3.9 plots average distance distributions, which have the potential to serve as an indicator for the success of the method, since, based on the previous analysis, the error is found to be low for low distances.

3.3 Semantics at a Larger Scale

Next, we analyze the potential of the resulting dataset. This analysis centers around the semantic signatures provided in Figures 3.10, 3.11, and 3.12, and is made available online.

3.3.1 Expressive Potential for Attributes

Figure 3.10 reveals the potential of the included fonts to represent different semantic attributes adequately. For a given attribute, the existence of high-scoring fonts entails a potential to convey that particular attribute effectively. In contrast, a narrower range of values limits this capability. Based on these considerations, we consider three categories of attributes.

High Potential. Attributes in this category are associated with fonts with a wide range of association scores, encompassing both very high (>0.8) and very low (<0.2) values. This is a high potential scenario because a well-chosen font can easily distinguish itself from the remaining fonts and may reflect the attribute more strongly. Based on the analysis in Figure 3.10, the attributes in this category are *angular*, *artistic*, *attention-grabbing*, *attractive*, *boring*, *complex*, *dramatic*, *happy*, *modern*, *playful*, *sloppy*, *strong*, and *thin*.

Moderate Potential. These attributes possess a high average value, which, at first glance, might be taken as implying a high potential. Yet, this also suggests a potential challenge in emphasizing the attribute more markedly. Still, creating a strong representation may be possible if fonts for other attributes (perhaps opposite attributes) exist in the same context. For this reason, we consider the following attributes as moderate potential ones: *calm*, *charming*, *formal*, *fresh*, *friendly*, *gentle*, *graceful*, *legible*, *sharp*, *soft*, and *warm*.

Low Potential. We consider the attributes in this category as having low potential due to an absence of fonts with very high values (>0.8) for them. Specifically, the attributes in this category are *bad*, *clumsy*, *delicate*, *disorderly*, *pretentious*, and *technical*. Despite being categorized as showing limited promise, these attributes might still prove informative as to which attributes to explore as potential candidates for the *moderate potential* category (e.g., opposites of these attributes).

3.3.2 Quality

We use the nearest neighbor distances to further assess the quality of the dataset. As discussed in the previous section, our algorithm uses the most similar four fonts to determine the values for a new font. The success of the algorithm increases when the average distance to these similar fonts decrease. For this reason, the distance value may be interpreted as a confidence value (inverse relationship), although the evaluations reveal that in some cases it is still possible to have a successful prediction with a high distance value.

Figure 3.11 depicts a filtered plot considering only fonts with a low distance value (lower than ~0.27) and thus in the highest confidence bracket. The interesting finding here is that many of the fonts at the high or low end of the range that determine the category of a given attribute remain. In other words, these are the fonts that possess

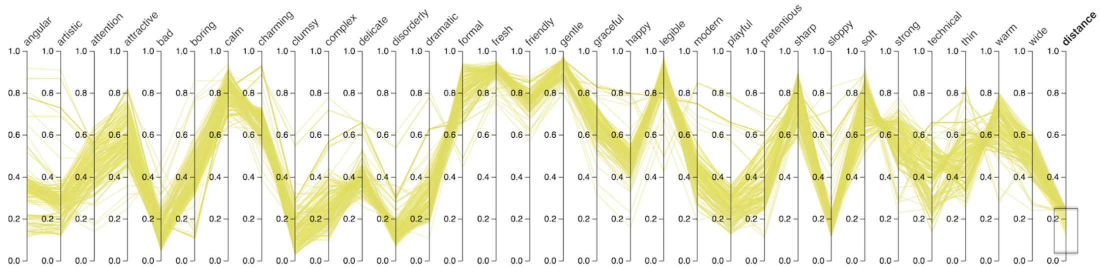


Figure 3.11: Semantic signatures of the lowest distance (highest confidence) fonts, as filtered from the visualization in Figure 3.10

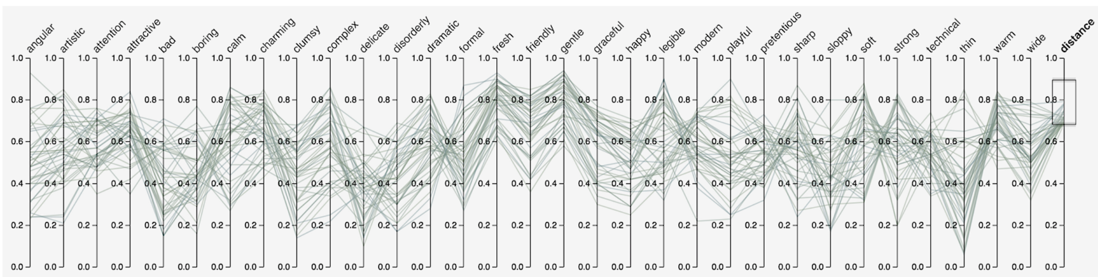


Figure 3.12: Semantic signatures of the highest distance (lowest confidence) fonts, as filtered from the visualization in Figure 3.10

an important relationship with the attribute, and, in the case of being at the high end, are the strongest candidates to be selected.

Figure 3.12 depicts a filtered plot considering only fonts with a high distance value (higher than ~0.7) and thus in the lowest confidence bracket. The results, again, appear favorable, as these fonts have scores that lie mostly closer to the middle of the respective ranges for each attribute. Thus, they have a smaller chance of being selected to represent those attributes. Nevertheless, this does not preclude the possibility of them in reality having values closer to the ones at the ends of these ranges, which would mean that we might be overlooking a font that could be a good candidate to represent an attribute.

Taken together, this suggests that the categorical organization of the attributes provided above is overall fairly reliable. In conjunction with the finding from Section 3.3 that most fonts have low-to-mid range distances, the fonts in our dataset, especially when picked from the ends of the ranges, tend to have very representative attribute values (see Figure 3.1).

3.4 Discussion

We now review and discuss the findings of this chapter, starting with the analysis in Section 3.1. Although there are some general trends in the data (such as high values for *gentle*), fonts appear to show characteristic biases. This is expected, as font categories are defined based on combinations of certain design metrics (contrast, x-height, etc.), which give rise to a particular perception with shared semantic characteristics. This is also confirmed by the scattered distribution of the font category *display*, since it is the only one among these categories with a very wide range of characteristics, complying with general design knowledge. Our results are also in line with the previous user study by Shaik et al. [22]. Both studies find *serif* to be *formal*, *monospace* to be *boring*, and *handwriting* to be *happy*.

Despite being able to reflect these category-based biases, the crowdsourced dataset is not large enough and the correlations not sufficiently clear to give rise to generalized metrics or models. To overcome this challenge, in Section 3.2, we use a k–NN approach. Our evaluation shows that this method has very low error rates for the font attribute prediction problem at hand. Another interesting point here is that the CNN embeddings are found to be a better similarity measure for attribute prediction compared to the typographical features.

Section 3.3 attempts to approximate the quality of the generated dataset, and makes predictions about the potential of the fonts to represent these semantic attributes. A point that should be noted is that all attributes have values growing away from the center (0.5). This is important because it shows that there is a high risk to unintentionally represent these attributes at different levels (high or low) if the font selection process does not consider these associations.

3.5 Conclusion

Starting with a crowdsourced dataset, we first analyzed the relationship between font categories/styles and semantic attributes, and reported a series of novel findings. We published an interactive online visualization that provides further insights from the dataset. Secondly, we induced a large-scale repository of semantic signatures for nearly 2,000 fonts, based on a weighted k-NN approach via a CNN embedding based similarity measure. Finally, we analyzed the resulting data to assess its quality, and provided an interactive visualization to allow for exploring it. We also characterized the potential of these fonts to represent different groups of semantic attributes. The work presented in this chapter was published in the proceedings of the 12th IEEE International Conference on Semantic Computing [61].

Chapter 4

FontLex: An Affective Font Lexicon

Towards the aim of supporting the development of font recommendation tools based on the *textual content* and the associated *affect* of the message, in this chapter, we develop methods to induce associations between words and fonts. We rely on word–emotion and font–emotion associations to connect words with fonts via their affective associations [62]. With these techniques, we induce a *typographical lexicon* called *FontLex*, which maps 6.7K words to a set of around 2K fonts. We further extend the resulting lexicon using the following approaches: 1) Using the dyads from Plutchik's Wheel of Emotion, in Section 4.3.1 we present a method to infer font vectors for complex emotions. 2) In Section 4.4, we extend the font set from 200 to around 2K using font embeddings following the approach from Chapter 3. 3) In Section 4.3.3, we demonstrate a sample poster design application which makes use of FontLex to achieve semantic font recommendation.

The rest of this chapter is organized as follows. Section 4.1 presents our method to predict emotion–font scores and evaluates it through a user study[1]. Section 4.2 presents our method to predict word–font scores using the previously obtained emotion–font scores, and evaluates it through a further user study. Subsequently, Section 4.3 describes the extensions we propose for the dataset to increase its accuracy and to expand its coverage to more words and fonts. In Section 6.7, we provide discussions on FontLex and its potential applications. Finally, Section 4.6 concludes the chapter and outlines plans for future work.

---

[1] All studies in this chapter received IRB approval.

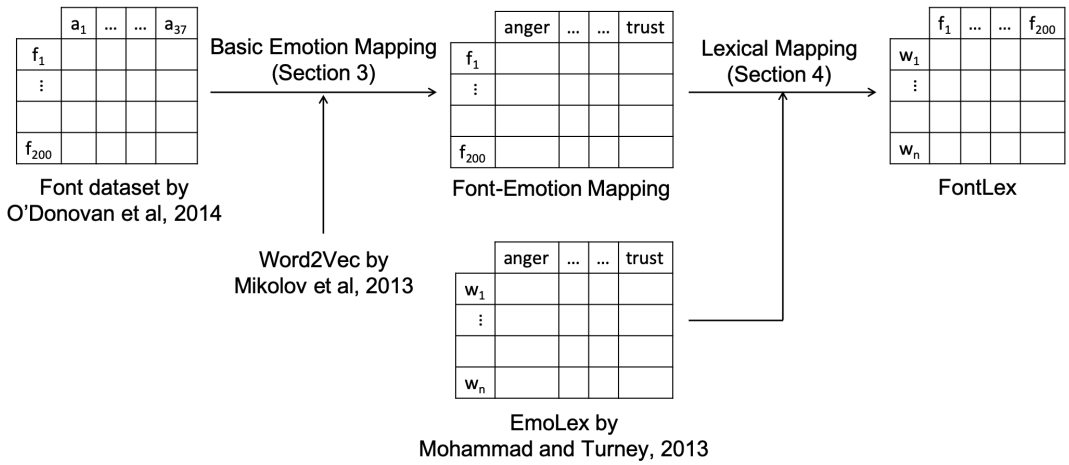


Figure 4.1: Overview of our approach to obtain FontLex, where f_i are fonts, a_i are font attributes, and w_i are words.

4.1 Basic Emotion Mapping

In this section, we describe the *Basic Emotion Mapping*, i.e., our method to obtain associations between fonts and emotional attributes. These will later, in the following section on the *Lexical Mapping* process, be used to obtain associations between fonts and words via their respective emotional associations. Figure 4.1 provides an overview of this process.

4.1.1 Method

Our method assumes as input a set of fonts $\mathcal{F}$ that are described in terms of a set of font attributes $\mathcal{A}$. For this, we rely on the same crowdsourced data [21] leveraged in Chapter 3, which for a given font $f \in \mathcal{F}$ provides scores in $[0, 100]$ for each attribute $a \in \mathcal{A}$. From this data, we derive $|\mathcal{F}|$-dimensional vectors $\vec{a} \in [0,1]^{|\mathcal{F}|}$ for each font attribute $a \in \mathcal{A}$. For this, we simply transform the dataset to consider the fonts for a given font attribute, normalizing scores to $[0, 1]$.

Then, to induce FontLex, we first generate $|\mathcal{F}|$-dimensional font vectors for a set of emotion attributes $\mathcal{E}$. Subsequently, using existing word–emotion associations, we will infer $|\mathcal{F}|$-dimensional font vectors for words such that each component of such a vector quantifies the strength of the association between a word and a font.

As the set of emotions $\mathcal{E}$, we consider the ten emotion attributes used in EmoLex [63]. Our first step is to map these $e \in \mathcal{E}$ to vectors $\vec{e} \in R^{|\mathcal{F}|}$ that characterize their association with fonts $f \in \mathcal{F}$ in our data.

index	emotion	1	2	3
1	anger	¬calm	clumsy	capitals
2	anticipation	fresh	formal	dramatic
3	disgust	clumsy	bad	sloppy
4	fear	bad	capitals	¬calm
5	joy	happy	playful	graceful
6	negative	bad	strong	sharp
7	positive	strong	¬bad	happy
8	sadness	¬happy	gentle	¬graceful
9	surprise	dramatic	happy	¬sharp
10	trust	strong	calm	¬bad

Table 4.1: Top three closest attributes for the basic emotions, where ¬ indicates attributes that are negated.

To achieve this, we proceed as follows. For each emotion $e \in \mathcal{E}$, we determine the $k = 3$ most similar font attributes $a \in \mathcal{A}$, as shown in Table 6.1. To decide on this value, we have carried out leave-one-out tests on the crowdsourced seed dataset [21]. Although the average overall success of the method in terms of the mean error was slightly higher for higher k than 3, we found that for $k = 3$ the most attributes attained their highest scores. Also considering the complexity of the negation decisions as will be described shortly, we opted to use the closest $k = 3$ neighbors.

We rely on word2vec [64] distances $d(e, a)$, using cosine distances on the standard word2vec Google News pretrained model[2], to determine similarity scores $\mathrm{sim}(e, a)$ between emotion names and font attribute names as below:

$$\mathrm{sim}(e, a_i) = \frac{1}{k-1} \frac{\sum_{\substack{j=1 \\ i \neq j}}^{k} d(e, a_j)}{\sum_{j=1}^{k} d(e, a_j)} \tag{4.1}$$

One aspect that needs to be addressed, however, is the widely known fact that distributional models of semantics tend to conflate synonyms with antonyms. Hence,

[2]https://code.google.com/archive/p/word2vec/

anger	anticipation	disgust	Fear	joy	NEGATIVE	positive	sadness	surprise	trust
ANGER	anticipation	disgust	FEAR	joy	NEGATIVE	positive	sadness	surprise	trust
anger	anticipation	disgust	FEAR	joy	negative	positive	sadness	surprise	trust

Figure 4.2: Emotion attributes rendered using the three most congruent fonts as predicted by our method. The renderings on the first line uses the fonts ranked 1st, the second line uses fonts ranked 2nd, and the third line uses fonts ranked 3rd.

anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	trust
anger	anticipation	disgust	fear	JOY	negative	positive	sadness	surprise	trust
anger	anticipation	disgust	fear	JOY	negative	positive	sadness	surprise	trust

Figure 4.3: Emotion attributes rendered using the neutral fonts as predicted by our method. The renderings on the first line use the fonts ranked 99th, the second line uses fonts ranked 100th, and the third line uses fonts ranked 101st.

we first define

$$\vec{\mu}(e,a) = \begin{cases} \vec{1} - \vec{a} & \text{if } a \text{ is assessed as an antonym of } e \\ \vec{a} & \text{otherwise,} \end{cases} \tag{4.2}$$

where $\vec{1}$ is an $|\mathcal{F}|$-dimensional vector of ones. Thus, for those words that are assessed as antonyms, we do not use the regular font vector $\vec{a}$, but instead consider an inverted vector, in which we subtract each value from the maximum value of 1. The assessment is performed manually. For relationships such as between *anger* and *calm*, determining antonym relationships was straightforward. However, for some more challenging decisions, such as *negative* and *sharp*, we evaluated both options and discussed the obtained results with a graphic designer before making the final decision. In Table 6.1, attributes labelled as antonyms are marked with a “¬” symbol.

To obtain font vectors $\vec{e}$ for emotions $e \in \mathcal{E}$, we compute

$$\vec{e} = \sum_{i=1}^{k} \mathrm{sim}(e, a_i)\, \vec{\mu}(e, a_i) \tag{4.3}$$

where the a_i are the k most similar attributes, as described above. Thus, the font vectors are a weighted average of the vectors for related attributes, after possibly inverting their respective vectors.

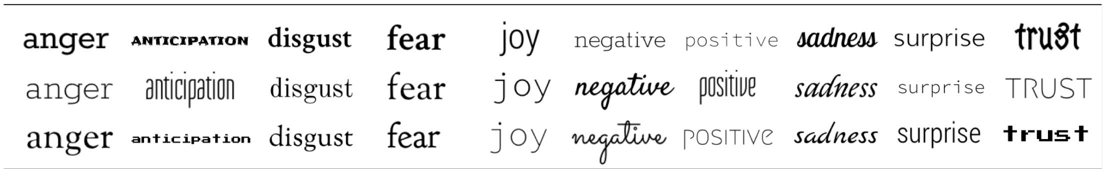


Figure 4.4: Emotion attributes using the three most incongruent fonts as predicted by our method. The renderings on the first line use the fonts ranked 198th, the second line uses fonts ranked 199th, and the third line uses fonts ranked 200th.

4.1.2 Results

Figure 4.2 depicts the top 3 fonts that are most strongly associated with the ten emotion attributes, whereas Figure 4.3 shows sample fonts that are predicted to be neutral in terms of the respective emotion. Figure 4.4 shows the three fonts for each emotion that are found to have the weakest associations. More specifically, the neutral fonts for emotion e are defined as those that are in the middle of the ranked font vector $\vec{e^r}$ of size n, namely $\vec{e^r}_i$ for $i \in \{\frac{n}{2}-1, \frac{n}{2}, \frac{n}{2}+1\}$, where $\vec{e^r}_1$ has the strongest association with the emotion, and $\vec{e^r}_n$ has the weakest association. In all figures, the emotion names are rendered using the corresponding fonts.

The fonts that are strongly associated with emotions share some special characteristics. For instance, for *joy*, we encounter handwriting-style typefaces, whereas for *disgust*, we find display fonts with salient stylization. It should also be noted that not all fonts that share these characteristics are strongly associated with these emotions, since the relationships between emotion attributes and font characteristics are not straightforward.

4.1.3 Evaluation

To assess the quality of the obtained emotion font score predictions, we carry out a user study.

User Study. For each of the ten emotion attributes, we generated four tasks with different random font choices. An example is given in Figure 4.5. Each task includes 5 fonts, two congruent fonts selected randomly among the top-scoring 10 fonts for that emotion, two incongruent fonts selected randomly among the lowest-scoring 10 fonts

Figure 4.5: An example task for *positive*. The second and fifth fonts are congruent, the third and fourth is incongruent and the first is neutral.

for that emotion, and one neutral font selected randomly among the ten fonts that are in the middle of the ranked list of fonts. In each task, the user is requested to select a single image that best reflects the semantics of the word. As described above, the available options include the same word presented using five different fonts.

Each task is carried out by 30 participants via Mechanical Turk, all from the United States, with at least 5,000 approved hits and an overall approval rating of 97% or more. We used counterbalancing, i.e., half of the users received the tasks in the reverse order from the other half. We also used three validation tasks, and eliminated results of three participants who incorrectly answered all three of them.

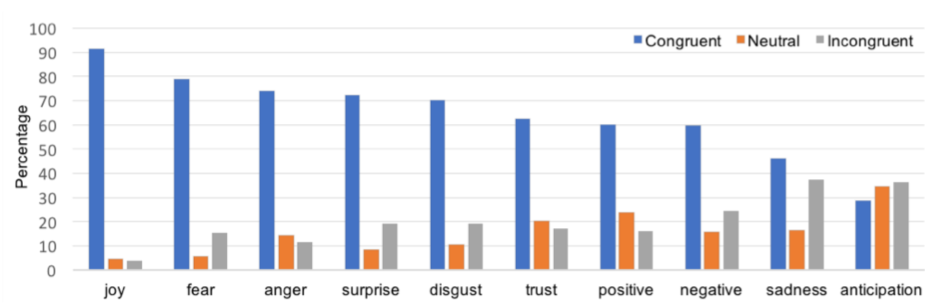


Figure 4.6: Results (in %) for the user study evaluating the obtained emotion-font associations. With uniformly random selections, congruent, neutral and incongruent options have a 40%, 20%, or 40% chance of being selected, respectively.

Emotion	$X^2(2)$	p
anger	52.803	$< .001$
anticipation	14.654	$< .001$
disgust	39.534	$< .001$
fear	65.519	$< .001$
joy	114.53	$< .001$
negative	17.069	$< .001$
positive	26.143	$< .001$
sadness	1.8413	.3983
surprise	44.726	$< .001$
trust	26.553	$< .001$

Table 4.2: Chi-square goodness of fit test results for the user study from Section 4.1.3.

Evaluation Results. Figure 4.6 summarizes the results of this user study. The *congruent* bars represent the percentages of selections in which the congruent fonts (those in the top 10 for that word) are preferred. Similarly, the *neutral* and *incongruent* bars represent the percentages of choices of neutral and incongruent fonts, respectively.

The average is 64% for congruent font preferences. Compared to the expected value of 40%, this shows a strong trend toward the fonts predicted to be congruent, hence validating our results in general. Similarly, the preferences for the fonts that are found to be incongruent by our method was much lower than the expected value, with an average of only 20%.

In Table 4.2, we provide statistical analysis for the user study results using chi-square goodness of fit test. This test compares observed sample distribution with the expected probability distribution to check whether there is a significant difference between the two. In this study, the expected probabilities are 0.4, 0.2, and 0.4 for the categories *congruent*, *neutral*, and *incongruent*, respectively. Results for all emotions except *anticipation* and *sadness* support our method based on a significance level set at 0.05. The emotion *sadness* has a p value above the significance level, and the emotion *anticipation* was significant but in an unexpected direction. This suggests that different emotions may differ in how saliently and uniquely they are associated with visual font characteristics (cf. Section *sec:discussion*).

4.2 Lexical Mapping

The next phase involves computing font vectors for words that reflect the degree of association between words and potential fonts. As shown earlier in Figure 4.1, we rely on the results of the Basic Emotion Mapping from Section 4.1 as our input, along with data from a word–emotion lexicon, to induce our FontLex resource.

4.2.1 Method

EmoLex [63] provides binary emotion association indicators between words and the emotion attributes $e \in \mathcal{E}$ listed in Table 6.1. There are 6,468 words with at least one

cab	certify	daughter	elegance	GUILTY	LIFELESS	loyalty	massacre	peaceful	resign
cab	certify	daughter	elegance	guilty	LIFELESS	loyalty	MASSACRE	peaceful	RESIGN
cab	certify	daughter	elegance	GUILTY	LIFELESS	loyalty	MASSACRE	peaceful	RESIGN

Figure 4.7: Selected words rendered using the three most congruent fonts as predicted by our method. The renderings on the first line uses the fonts ranked 1st, the second line uses fonts ranked 2nd and the third line uses fonts ranked 3rd.

cab	certify	daughter	elegance	guilty	lifeless	loyalty	massacre	peaceful	resign
cab	certify	daughter	elegance	guilty	lifeless	loyalty	massacre	peaceful	resign
cab	certify	daughter	elegance	guilty	lifeless	loyalty	massacre	peaceful	resign

Figure 4.8: Selected words rendered using the three fonts from the middle of the ranked list as predicted by our method. The renderings on the first line uses the fonts ranked 99th, the second line uses fonts ranked 100th and the third line uses fonts ranked 101st.

cab	certify	daughter	elegance	guilty	lifeless	loyalty	massacre	peaceful	resign
cab	CERTIFY	daughter	elegance	guilty	lifeless	loyalty	massacre	peaceful	resign
CAB	certify	daughter	elegance	guilty	lifeless	LOYALTY	massacre	peaceful	resign

Figure 4.9: Selected words rendered using the three most incongruent fonts as predicted by our method. The renderings on the first line uses the fonts ranked 198th, the second line uses fonts ranked 199th and the third line uses fonts ranked 200th.

emotion association in their data. For words w in this set, we consider their data as providing vectors $\vec{w}_{\mathrm{E}} \in [0,1]^{|\mathcal{E}|}$.

To generate a font vector $\vec{w}_{\mathrm{F}}$ for a word w, we compute

$$\vec{w}_{\mathrm{F}} = \frac{1}{\|\vec{w}_{\mathrm{E}}\|_1} \mathbf{M}_{\mathrm{E}} \vec{w}_{\mathrm{E}} \tag{4.4}$$

where $\|\vec{w}_{\mathrm{E}}\|_1$ denotes the ℓ_1 norm of $\vec{w}_{\mathrm{E}}$ and $\mathbf{M}_{\mathrm{E}} = [\vec{e}_1 \ldots \vec{e}_{|E|}]$, i.e., a matrix with columns that capture the font vectors for the emotions $e \in \mathcal{E}$ (in the same order as captured in $\vec{w}_{\mathrm{E}}$).

4.2.2 Results

Figure 4.7 shows the top three congruent fonts associated with ten sample words, Figure 4.8 shows sample fonts that are predicted to be neutral for the respective words, and Figure 4.9 shows the most incongruent three fonts for the same words. In all images, the words are rendered using the corresponding fonts. These words are among those used in the evaluation user study in the following section.

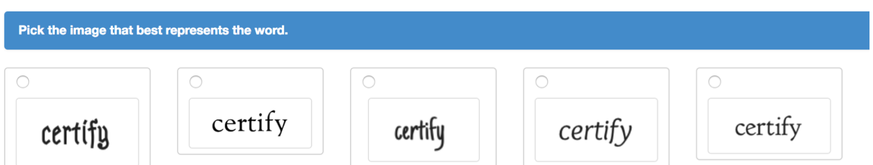


Figure 4.10: An example task for the word *certify*. The second and fifth fonts are congruent, the first and third is incongruent, and the fourth is neutral.

4.2.3 Evaluation

We evaluate the dataset through a user study. In the following, we provide details on the design and the results of this study.

User Study. For our study, we consider 25 words randomly selected from the set of words with at least one salient font association. For this purpose, we consider any of the 3,882 words that have a score of 0.75 or higher in any of the components of their respective font vectors. For each of the random 25 words, we generated two tasks with different random font choices. We have reduced the number of tasks to two, compared to the four tasks used in the previous section, to keep the total number of tasks reasonable for each participant.

An example task for the word *certify* is given in Figure 4.10. Each task includes 5 fonts, two congruent fonts selected randomly among the top-scoring 5 fonts for that word, two incongruent fonts selected randomly among the lowest-scoring 5 fonts for that word, and one neutral font selected randomly among the three fonts that are in the middle of the ranked list of fonts for the word. The decision to use 5 fonts as opposed to 10 is again based on considerations regarding the workload per user.

Each task involves a user being requested to select the image that best represents the word. As described above, the available options include the same word presented using five different fonts. Each task is carried out by 30 participants in Mechanical Turk, all from the United States, with at least 5,000 approved hits and an overall approval rating of 97% or more. We used counterbalancing and eliminated results of one participant that accidentally completed both of the original and reversed task sessions. We have also used three validation tasks, and eliminated results of one participant that incorrectly answered both of the two validation tasks.

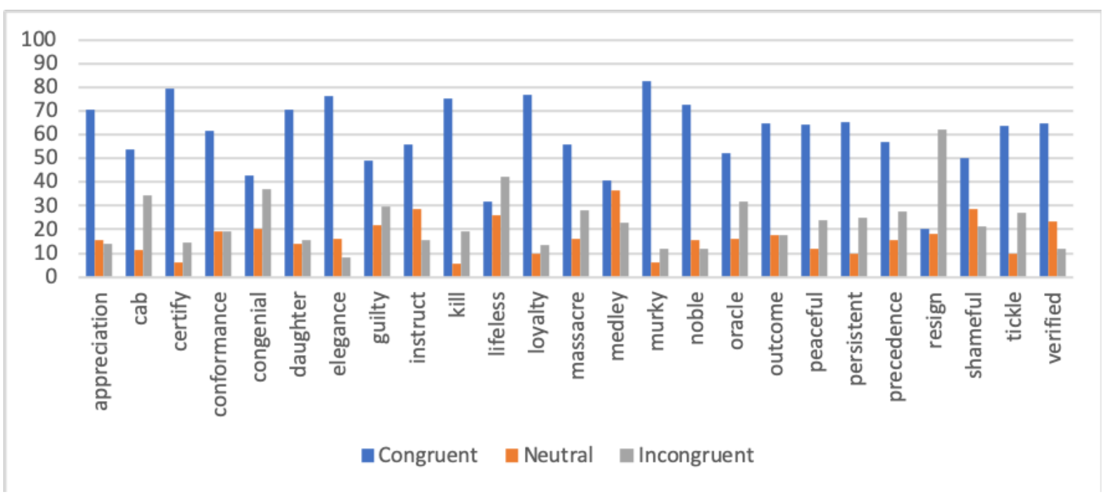


Figure 4.11: Results (in %) for the user study evaluating the obtained word-font associations.

	AG	AN	D	F	J	N	P	SA	SU	T
appreciation	0	0	0	0	1	0	1	0	0	1
cab	0	0	0	0	0	0	1	0	0	0
certify	0	0	0	0	0	0	0	0	0	1
conformance	0	0	0	0	0	0	1	0	0	0
congenial	0	0	0	0	0	0	1	0	0	0
daughter	0	0	0	0	1	0	1	0	0	0
elegance	0	1	0	0	1	0	1	0	0	1
guiilty	1	0	0	0	0	1	0	1	0	0
instruct	0	0	0	0	0	0	1	0	0	1
kill	0	0	0	1	0	1	0	1	0	0
lifeless	0	0	0	1	0	1	0	1	0	0
loyalty	0	0	0	0	0	0	1	0	0	1
massacre	1	0	1	1	0	1	0	1	0	0
medley	0	0	0	0	0	0	1	0	0	0
murky	0	0	1	0	0	1	0	1	0	0
noble	0	0	0	0	0	0	1	0	0	1
oracle	0	1	0	0	0	0	1	0	0	1
outcome	0	0	0	0	0	0	1	0	0	0
peaceful	0	1	0	0	1	0	1	0	1	1
persistent	0	0	0	0	0	0	1	0	0	0
precedence	0	0	0	0	0	0	1	0	0	1
resign	1	0	1	1	0	1	0	1	0	0
shameful	0	0	0	0	0	1	0	1	0	0
tickle	0	1	0	0	1	0	1	0	1	1
verified	0	0	0	0	0	0	1	0	0	1

Table 4.4: Emotion associations for the words from the user study. (AG: Anger, AN: Anticipation, D: Disgust, F: Fear, J: Joy, N: Negative, P: Positive, SA: Sadness, SU: Surprise, T: Trust)

Evaluation Results. Figure 4.11 summarizes the evaluation results for the 25 randomly selected words as described above. The *congruent* bars represent the percentages of selections in which the congruent fonts (those in the top 5 for that word) are preferred. Similarly, the *neutral* and *incongruent* bars show the percentages of choices of neutral and incongruent fonts, respectively.

The average is 59% for congruent font preferences, which shows that the consensus between our data and the users were strong. The strongest preference is obtained for the word *murky*, with a value of 82%, whereas the lowest is for the word *resign* with 20%. Similarly, the average for the incongruent preferences was only 23%, bearing further witness to the quality of the results. Only two out of twenty-five words, namely *lifeless* and *resign*, received congruent preferences that are less than the expected value of 40%. Such results are expected, given that different words may differ in the strength and uniqueness of their associations (cf. Section 6.7).

Table 4.4 displays the corresponding emotions for the words used in the evaluation, allowing us to analyze the relationship between the success of the two datasets. In some cases, words associated with the same set of emotions obtained similar user ratings, such as *instruct*, *noble*, *precedence*, and *verified*. Whereas in some cases, words with the same emotion set obtained quite divergent ratings: *massacre* and *resign*.

4.3 Extensions to FontLex

In this section, we present methods to extend the dataset and increase its accuracy.

4.3.1 Complex Emotion Mapping

We propose using Plutchik's Wheel of Emotion [27] to infer font scores for complex emotions (e.g., hope). Plutchik's theory suggests that complex emotions are indeed combinations of basic ones, referred to as *dyads*. For instance, the theory posits that *hope* is a superposition of the more basic emotions *anticipation* and *trust*. Relying on the font vectors $\vec{e}$ obtained in Section 4.1, we compute font vectors $\vec{c}$ for complex emotions $c \in \mathcal{C}$ as

$$\vec{c} = \frac{1}{2}(\vec{e}_i + \vec{e}_(j)),$$

where the e_i and e_j are the underlying basic emotions for c, and i and j are their indices from Table 6.1.

Figure 4.12 provides dyads for all $c \in \mathcal{C}$, while rendering the words for each basic and complex emotion with the most congruent font as inferred by our study. As an example, the first entry suggests that *anticipation* and *joy* together evoke the feeling of *optimism*. The font determined as most congruent for *optimism* appears to combine visual characteristics of both its underlying basic emotions, namely *anticipation* and *joy*.

anticipation	+	joy	→	optimisim		surprise	+	sadness	→	DISAPPROVAL
anticipation	+	fear	→	ANXIETY		surprise	+	disgust	→	unbelief
anticipation	+	trust	→	hope		surprise	+	anger	→	OUTRAGE
joy	+	trust	→	love		sadness	+	disgust	→	remorse
joy	+	fear	→	GUILT		sadness	+	anger	→	ENVY
joy	+	surprise	→	delight		sadness	+	anticipation	→	pessimism
trust	+	fear	→	SUBMISSION		disgust	+	anger	→	contempt
trust	+	surprise	→	curiosity		disgust	+	anticipation	→	cynic
trust	+	sadness	→	sentimentality		disgust	+	joy	→	morbid
fear	+	surprise	→	AWE		anger	+	anticipation	→	aggressive
fear	+	sadness	→	DESPAIR		anger	+	joy	→	PRIDE
fear	+	disgust	→	shame		anger	+	trust	→	DOMINANCE

Figure 4.12: Emotion combinations (dyads) from Plutchik's Wheel of Emotion, rendered using the congruent fonts as determined by our study.

	basic emotions	sentiment
aggressive	anger, fear	negative
anxiety	anger, anticipation, fear, sadness	negative
awe	–	–
contempt	anger, disgust, fear	negative
curiosity	anticipation, surprise	positive
cynic	–	–
delight	anticipation, joy	positive
despair	anger, disgust, fear, sadness	negative
disapproval	sadness	negative
dominance	–	–
envious	–	negative
guilt	disgust, sadness	negative
hope	anticipation, joy, surprise, trust	positive
love	joy	positive
morbid	sadness	negative
optimisim	anticipation, joy, surprise, trust	positive
outrage	anger, disgust	negative
pessimism	anger, fear, sadness	negative
pride	joy	positive
remorse	sadness	negative
sentimentality	–	positive
shame	disgust, fear, sadness	negative
submission	–	–
unbelief	–	negative

Table 4.5: Basic emotion and sentiment associations of complex emotions as suggested by EmoLex.

The obtained font scores for complex emotions serve two purposes. The first is that, similar to the basic emotions, they could be used as seed information to infer higher-quality font vectors for arbitrary words. A second purpose is to override font vectors for the complex emotion words in FontLex, potentially improving its accuracy.

To explore this more, we assess how the complex emotions described by Plutchik's Theory are annotated in EmoLex [63]. For these complex emotion words, Table 4.5 lists the corresponding basic emotions as given by EmoLex. The most notable problem is that 7 of the 24 complex emotions are not associated with any basic emotions. This might stem from issues of ambiguity, e.g., for *submission*. However, 3 of these 7 words with missing emotions are actually assigned a sentiment, which reduces the likelihood of such issues for these entries. Indeed, none of these associations are exact matches to their corresponding dyads. In total, 11 complex emotions have emotions assigned but miss at least one of the dyad emotions (e.g., love, guilt). Out of these, five complex emotions have only one basic emotion associated. In addition, 6 complex emotions

boring	DOMINANCE	joy	WIDE	sharp
deadening	ascendance	pleasure	broad	crisp
slow	say-so	gladden	extensive	acute
irksome	AUTHORITY	rejoice	FULL	needlelike
dull	CONTROL	joyousness	spacious	knife

Figure 4.13: Examples of synonyms retrieved form WordNet for the attributes *boring*, *dominance*, *joy*, *wide* and *sharp*; rendered using congruent fonts.

have the two emotions from the dyad defined by Plutchik, but also additional ones (e.g., optimism, hope).

Overall, we conclude that EmoLex is incomplete in its description of complex emotion words, and that relying on Plutchik's theory can yield better font associations.

4.3.2 Semantic Relationships

We extend the dataset and increase its accuracy by accounting for semantic relationships given by WordNet [65]. For all attribute words in $\mathcal{A} \cup \mathcal{E} \cup \mathcal{C}$, in total 71 attributes (37 original font attributes from [21], 10 basic emotion attributes as computed in Section 4.1, and 24 complex emotion attributes as computed in Section 4.3.1), we gather synonyms. For the original font attributes, we gather the set of words that share a common synset with the attribute names (such as the words *deadening*, *dull*, *ho-hum*, *irksome*, *slow*, *tedious*, *tiresome* and *wearisome* for the font attribute *boring*). We then go through this list manually to exclude any synonyms with an irrelevant meaning (such as the word *building complex* for the font attribute *complex*). For the basic and complex emotion attributes, we pick the sense describing an emotion, and then use the synonyms from these synsets. These synonyms are assigned the font vectors of the corresponding words in $\mathcal{A} \cup \mathcal{E} \cup \mathcal{C}$. This results in 464 additional word-font assignments, 166 of which override the ones from the methods in Sections 4.1 and 4.2 While small in number, these provide for particularly salient associations (examples provided in Figure 4.13).

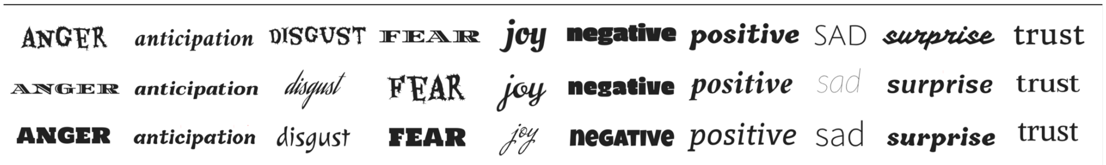


Figure 4.14: Basic emotions rendered using the three most congruent fonts from the extended font set (excluding 200 fonts from the original dataset) as predicted by our method.

4.3.3 More Fonts

Our study relies on the data from [21], which connects 200 fonts with 37 attributes. In the previous sections, we extend its attribute set and connect it with the words from EmoLex, keeping the font set the same. We now proceed to extend our lexicon to use further fonts following the method proposed by **?**].

Our goal is to predict font vectors $\vec{f'}$ for fonts $f' \notin \mathcal{F}$. To achieve this, we use weighted k-nearest neighbors (k–NN) regression using 4 neighbors. The weighted k–NN approach generates weights using the following equation.

$$w_i = \frac{1}{3}\frac{\sum\limits_{\substack{j=1 \\ i \neq j}}^{4} d(f', f_j)}{\sum\limits_{j=1}^{4} d(f', f_j)} \tag{4.5}$$

The distance between two fonts, denoted as $d(f_i, f_j)$ is computed using Convolutional Neural Network (CNN) embeddings from **(author?)** [55]. For each font, an image is generated rendering the letters (L,a,s,e,g,d,h,u,m,H,l,o,i,v) on a grid. These images are processed by the CNN and the obtained representations can be regarded as visual font embeddings. The visual distance between two fonts can then be computed as the Euclidean distance of their visual font embeddings.

Subsequently, the weighted values are generated as follows:

$$\vec{f} = \sum_{i=1}^{4} w_i \vec{f_i} \tag{4.6}$$

Using the above approach and aforementioned embeddings, we extend our dataset from 200 to 1,922 fonts, while each font vector include scores for every word in $\mathcal{A} \cup \mathcal{E} \cup \mathcal{C}$.

Figure 4.15: A word cloud of complex emotions rendered using fonts from the extended font set (excluding 200 fonts from the original dataset) that are inferred to be congruent by our method.

Figure 4.14 presents basic emotions using the most congruent three attributes from the extended dataset. We exclude the fonts from the original dataset to be able to portray results from just the extension.

Finally, the word cloud in Figure 4.15 provides complex emotions rendered with corresponding high-scoring fonts from the extended font set.

4.4 Application Example

In this section, we introduce a proof of concept *Poster Design* application for which FontLex could prove useful. In this example, the tool provides two types of support. In the first scenario, the tool recommends a font for a poster based on the words it includes. We compute a font vector $\vec{p}$ for a poster P as follows:

$$\vec{p} = \frac{1}{n}\sum_{i=1}^{n}\vec{w}_i \tag{4.7}$$

where $\vec{w}_i$ are the font vectors for the words w_i in poster P. We omit the words for which no font vector is found in FontLex. A sample is provided in Figure 4.16. In this example, the image on the left shows the poster with the default font, whereas the image on the right makes use of the recommended font (the font with the highest score in $\vec{p}$).

In the second scenario, each word is assigned a different font. For proof of concept purposes, we focus on semantic congruence, and ignore other important design concerns such as the harmony of different fonts. For a word w_i in P, the fonts with the highest

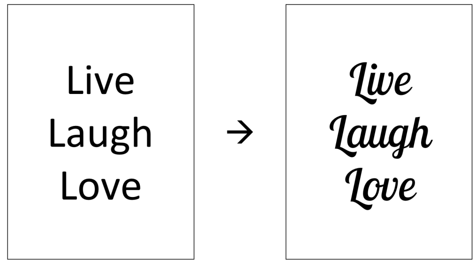


Figure 4.16: A poster design example where the congruent font recommendations for the poster is generated using FontLex.

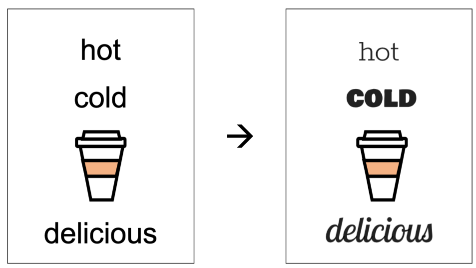


Figure 4.17: A poster design example where a congruent font is individually recommended for each word using FontLex.

values in $\vec{w}_i$ are considered as candidate fonts and one of them is selected randomly. Figure 4.17 provides an example, in which different fonts are assigned to the words in the poster. The poster on the left uses the default font, whereas the poster on the right makes use of the recommended fonts.

4.5 Discussion

In this section, we discuss our results and potential applications of FontLex and of dyads as a means of inferring complex emotions.

4.5.1 Results

We have introduced two datasets that connect emotions and words with fonts in terms of real-valued scores. Besides showing strong support for the datasets, the user evaluations also revealed that the performance varies for different emotions and words. Below, we discuss the potential sources for these differences.

For the emotion–font dataset, one reason for the differences between results could be the varying potential of fonts to represent or evoke different emotions. This could be observed in the results for *anticipation*, for which determining a font type may prove difficult even for an experienced graphic designer. It is also observed that emotions with higher arousal, namely *anger*, *disgust*, *fear*, *joy*, and *surprise*, received higher congruent user preferences compared to other emotions, which may be a direction that merits further analysis.

The second reason may be a lack of appropriate similar attributes in the crowd-sourced seed dataset. Looking at Table 6.1, it could be argued that *joy* has semantically close neighbors in the dataset, whereas this is not the case for *anticipation*.

For the word–font dataset, assessing the underlying emotion connections in Table 4.5 may shed some light on the differences. Recalling that the lowest performing emotion–font scores are for *anticipation* and *sadness*, one might expect that words associated with these emotions are prone to showing fewer user preferences that are congruent. The words associated with *anticipation*, namely *elegance*, *oracle*, *peaceful*, and *tickle*, do not seem to possess the same difficulty, as the lowest preference for these words is 52% (for *outcome*), which shows a strong preference.

On the other hand, among the words associated with *sadness*, the words *lifeless* and *resign* do not show such strong preferences. One might conjecture that this stems from low-performing emotion–font associations. However, looking at this in more detail, we find that *kill* and *massacre* have the same underlying emotion associations as *lifeless* and *resign*, respectively. The fact that the fonts for *kill* and *massacre* received strong support from users suggests that the word–emotion associations might have played a role. Some words may have inaccurate or missing emotion associations, while other words may have weaker emotional associations than others, which is not reflected in the binary scheme used by EmoLex. Using a dataset with real-valued scores instead of binary associations might help to capture the latter case.

Fortunately, overall, both datasets have received strong support from users, with around 60% and 64% of the average user preferences towards the fonts found to be congruent by our datasets. Only for two words out of twenty-five, incongruent fonts are preferred more frequently than chance would predict, i.e., $\frac{2}{5} = 40\%$. In contrast, for 23 words, congruent fonts are preferred more frequently than chance would predict. Despite the subjective nature of font preferences and associations, we observe that there is a clear correspondence between the fonts chosen by our method and those assessed as appropriate by the human participants.

4.5.2 Application Areas

The main use cases we foresee for FontLex are font search and font recommendation.

Semantic Font Search. Currently, content creation tools that heavily rely on text, such as word processors or graphic design tools, use traditional search methods for font search. [21] propose semantic attribute based font search as a step towards sufficient user support. We believe FontLex can help taking semantic search one step further by providing search using any keyword instead of a predefined small set of attributes. This could help users make use of a large number of fonts which is otherwise hard to achieve. Its flexibility would also allow users to be more creative.

Font Recommendation. Font–emotion mappings and FontLex could be utilized to enable semantic font recommendation, and we demonstrate such usage in Section 4.4. In addition to the support described in the example, FontLex could be utilized to provide more advanced support using some of its attributes (e.g., *legible* for readability, *artistic* for aesthetics) as filtering options. For instance, in our poster design application example, fonts could be filtered to pick only the *display* ones.

4.5.3 Emotion Combinations (Dyads)

In this study, we use combinations of basic emotions to calculate scores for the complex emotions based on the dyads provided by [27]. Based on our qualitative analysis, it is a powerful method to infer complex emotions, which is otherwise a challenging task. To the best of our knowledge, dyads have not been utilized before to infer complex emotions of content. Thus, a similar approach could be applied to other domains, such as for text and image.

4.5.4 Personal and Demographic Differences

Fonts are strongly tied to cultural elements, and hence may bear associations with various concepts, such as historical epochs, brands, or even music genres. Although we do not explicitly explore these connections, we believe the seed dataset that we rely on [21] accounts for such associations implicity, as it is a crowdsourced dataset and the emotional ratings that users provide are affected by such connections. At the same time, there are also differences between users, especially based on their demographics, such as culture, gender, and age. These personal and demographic differences of font semantics remain to be explored.

4.6 Conclusions

Our study aims to support the development of font recommendation tools. Following this aim, we have created FontLex, a dataset that maps 6.7K words to 1,922 fonts. These derive mainly from the affective associations between words and fonts. The work presented in this chapter was published in the proceedings of the 11$^{\text{th}}$ International Conference on Language Resources and Evaluation [66] and in Language Resources and Evaluation [67].

Chapter 5

Multimodal Font Discovery

Given that thousands of fonts are now freely available online, selecting among them is typically carried out via associated semantic and typographic tags. However, supporting users in deciding which fonts to pick is challenging when this is based only on such tagging. The ability of users to explore the different fonts is limited both by the incompleteness of the tagging and the limited tag inventory. If the tag inventory grows, the risk of missing tags for fonts increases. Even in an ideal scenario with a large tag inventory and in the absence of any missing tag associations, users would still suffer from the large number of fonts they would need to browse through to eventually find the ideal font for their use case.

In a recent user study, Wu et al. interviewed design practitioners regarding their font selection process and the challenges they faced [30]. Unsurprisingly, one of the main difficulties reported by the participants was identifying fonts that match a particular semantic profile. One participant reported this as follows:

> "When I'm looking for a particular font, [I] know what feeling [I] want the font to have. But I just spent so much time browsing and browsing, and still couldn't find the one."

This suggests the need for systems that support a more open-ended form of font discovery, allowing users to search for arbitrary attribute query words, including ones that are not present as tags in the data at all.

In the same user study, the participants also expressed their desire to *slightly modify* fonts that otherwise partially fulfilled their needs but were "just a little bit off" [30].

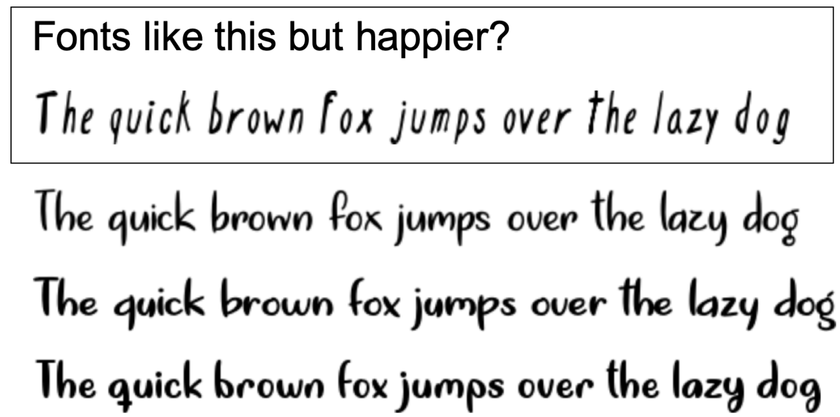


Figure 5.1: An example font search using the proposed multimodal querying strategy.

They further emphasized the need for *unique* fonts, so as to avoid very popular fonts and better differentiate their design product from those of competitors.

In this chapter, we propose a new multimodal font discovery method in which users provide a reference font that is visually similar to what they are seeking but only partially fulfills their needs, along with the changes they would like to obtain to get closer to their ideal font. In our proposed method, the changes are specified using keywords. Focusing on a similar problem, Ishibashi1 et al. [50] allow users to specify the stylistic changes they want to see by modifying the sketch of the reference font. Laenen et al. [68], on the other hand, propose a form of multimodal search that involves specifying keywords together with the reference item, but for the fashion domain.

Figure 6.1 shows an example of our proposed approach. If the user likes the style of a certain font, but needs a *happier* version of it, they can provide that font as a reference and indicate the change(s) they wish to have.

Using this mechanism, the users not only satisfy their need to *slightly modify* a font, but also have the ability to explore niche sections of the available font inventory to find a *unique* font, without spending their effort on reviewing fonts that are far from what they need. We enable this form of search strategy by embedding fonts and words into a joint cross–modal representation space, enabling the use of multimodal vector arithmetic.

Mountains of Christmas

Figure 5.2: Sample font "Mountains of Christmas" with the tags: *serif*, *christmas*, *bouncy*, *staggered*, *curly*, *cute*, *playful*, *casual*, *warm*, *fun*, *handwritten*, *text*, *google web*.

The above technique not only enables novel font discovery methods, but also helps overcome other semantic challenges, specifically, the challenges of limited tag inventories and of missing font–tag connections. Users obtain access to the entire vocabulary that the language (in our case English) provides, and a font need not be tagged with the specific words that users associate them with, since the method is able to infer such connections.

The rest of the chapter is organized as follows. Section 5.1 describes our data acquisition process to procure a large tagged font collection. In Section 5.2, we introduce our method to induce cross–modal vector representations. Section 5.3 then presents how we can use our method to search for fonts based on an arbitrary desired attribute, while Section 5.4 describes how we can invoke it to estimate font similarity, so as to find fonts based on a reference font. These are the two key building blocks of our multimodal search strategy, which is presented in Section 5.5. We conclude the chapter in Section 6.8 with a brief summary and discussion of our results.

5.1 Font Tagging Data

Our study assumes a large collection of fonts along with substantial (yet incomplete) social tagging. In the following, we describe how we procure such a dataset.

5.1.1 Data Crawling

We collected font–tag associations from www.1001fonts.com, a website that catalogs font files along with user-assigned tags. In Figure 5.2, a sample font name is shown together with its associated tags. As for most such Web resources, font families are tagged as a whole, e.g., the *italic* or **bold** versions of a typeface are not tagged separately. Similar to previous work [26], we adopt the "regular" version of a font family for

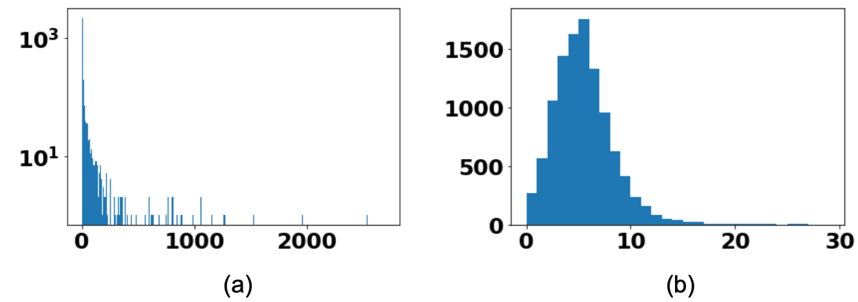


Figure 5.3: Histograms analyzing tag frequencies. (a) Distribution of tag frequencies across the entire dataset. (b) Distribution of tag counts per individual fonts.

Figure 5.4: Tags *digital*, *hairline*, *bouncy*, *playful*, *futuristic*, *halloween*, *graffiti*, *handwritten*, and *children* rendered using examples of fonts tagged accordingly in the dataset.

use in our dataset. Unlike previous studies, however, we apply a series of data cleaning steps to reduce the noise to the extent possible.

5.1.2 Data Cleaning

We filter out irrelevant fonts and tags as an attempt to clean otherwise noisy Web data.

Filtering Out Fonts. Dingbat fonts are fonts that consist entirely of symbols instead of alphabetical or numerical characters. They are used for decorative or symbolic purposes. As they are not relevant in rendering text, we discard all fonts assigned the *dingbat* tag in the data, which accounts for around 600 fonts.

Figure 5.5: Sample emotion-expressing attributes rendered using fonts tagged accordingly in the dataset.

Filtering Out Attributes. As we are interested in tags that describe semantic attributes of fonts and enable font discovery along such attributes (e.g., "happier"), we eliminate around 100 tags that merely denote font families (e.g., *serif*, *sans-serif*, *slab serif*) or other types of information (e.g., *google web*, *10pt*, *12pt*) that are not directly related to font semantics. We also eliminate a few tags that are not in English. We retain typographical tags that have the potential to provide semantic connections, such as *wide*, *handwritten*, *gothic*, *poster*, and *outlined*.

As a concrete example, for the font given in Figure 5.2, the tags *serif*, *text*, *google web*, and *medium* are eliminated, leaving the font with the tags *christmas*, *bouncy*, *staggered*, *curly*, *cute*, *playful*, *casual*, *warm*, *fun*, *handwritten*, and *light*.

5.1.3 Dataset Summary

After the above filtering, the resulting dataset contains around 10.4K fonts, 2.6K tags, and 54K font–tag assignments, with an average of 5 tags per font. Figure 5.3 shows the distributions of (a) overall tag frequencies and (b) tag counts per font. Most tags are used to tag fewer than a hundred fonts, and most fonts have fewer than 10 tags. Figure 5.4 displays three font examples for nine selected tags from the dataset, aiming to give a feeling of the range of the semantic connections. Figure 5.5 provides examples of fonts for sample emotion-expressing attributes. Figure 5.8 in Section 5.3 also provides examples of fonts for the ten most frequent attributes.

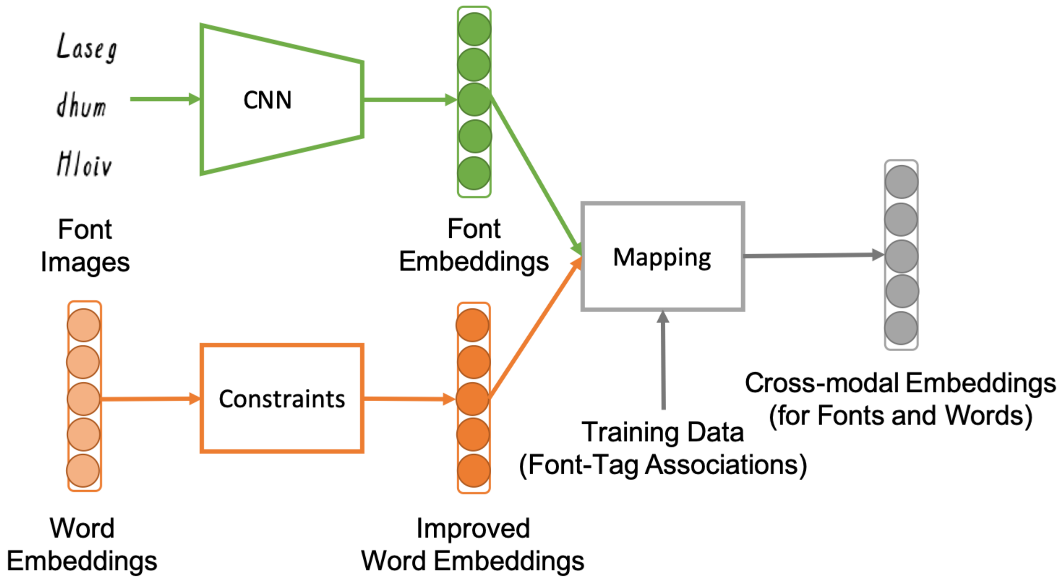


Figure 5.6: Overview of the proposed cross–modal representation induction method.

5.2 Cross–Modal Representation Learning

In order to facilitate identifying fonts that are similar to a given input font but differ along a particular attribute ("like this but happier"), we induce a cross–modal vector representation space. This not only allows us to jointly embed both fonts and query words in a single vector space, but also allows us to conduct vector arithmetic to locate fonts that better match a given semantic profile.

Our vector space induction method is summarized in Figure 5.6. We induce font embeddings using a deep convolutional neural network, and induce word embeddings by modifying pretrained distributed word embeddings to better satisfy antonymy and synonymy constraints. The final step is to connect the aforementioned font and word embeddings in a single cross–modal vector space.

Our method assumes as input a set $\mathcal{F}$ of fonts, which are associated with a set $\mathcal{A}$ of font attributes via a Boolean font–attribute matrix $\mathbf{M} \in \{0,1\}^{|\mathcal{F}|\times|\mathcal{A}|}$ based on the data described in Section 5.1.

5.2.1 Font Embedding Induction

Our first goal is obtain a font embedding matrix $\mathbf{F} \in R^{|\mathcal{F}| \times d}$ that in its rows provides a d-dimensional vector representation $\mathbf{v}_f \in R^d$ for each font $f \in \mathcal{F}$.

These vector representations are expected to reflect visual similarity, i.e., fonts f, f' that are visually similar ought to have similar vectors $\mathbf{v}_f$, $\mathbf{v}_{f'}$. To achieve this, for each font $f \in \mathcal{F}$, we generate an image rendering a fixed set of 14 different letters from the alphabet using that font so as to demonstrate its visual characteristics.

We then feed these images into a deep convolutional neural network with residual connections, specifically a ResNet-18 [69] model pre-trained on ImageNet [70]. For each font, we extract the resulting 512-dimensional latent representation from the average pooling layer of the model.

Finally, for dimensionality reduction to $d = 300$ dimensions, we apply Principal Component Analysis (PCA) and project every latent font representation into the space spanned by the first d principal components in order to obtain the desired matrix $\mathbf{F}$ with d-dimensional vectors $\mathbf{v}_f \in R^d$ for fonts $f \in \mathcal{F}$.

5.2.2 Word Embedding Induction

Our next goal is to induce vector representations of tags. We start out with the widely used 300-dimensional word2vec vectors pretrained on a large Google News dataset [71], which provides a word embedding matrix $\mathbf{W} \in R^{|\mathcal{V}| \times d}$ for a large vocabulary $\mathcal{V}$ of English words. The vectors are based on contextual information and the corresponding vector similarities reflect distributional similarity.

However, distributional similarity in general and word2vec word vectors in particular tend to give similar representations to words with opposite meaning such as *formal* and *informal* [72]. To alleviate this issue, we apply the Counter-fitting algorithm [73] to transform the original word embedding matrix $\mathbf{W}$ into a new embedding matrix $\mathbf{W}'$ subject to antonymy constraints A and synonymy constraints S. The algorithm minimizes the loss function

$$\ell(\mathbf{W}, \mathbf{W}') = \sum_{(u,w)\in A} 1 - d(\mathbf{v}'_u, \mathbf{v}'_w) + \sum_{(u,w)\in S} d(\mathbf{v}'_u, \mathbf{v}'_w) + \sum_{w\in\mathcal{V}} \sum_{u\in N(w)} \max(\mathbf{0}, d(\mathbf{v}'_u, \mathbf{v}'_w) - d(\mathbf{v}_u, \mathbf{v}_w)), \tag{5.1a}$$

where the notation $\mathbf{v}_w$ denotes the vector for w in $\mathbf{W}$, $\mathbf{v}'_w$ denotes the vector for w in $\mathbf{W}'$, and $N(w)$ denotes the set of nearest neighbors of w in $\mathbf{W}'$ with cosine similarity $\geq \tau = 0.8$. For the setting of τ as well as the constraint sets A and S, which are extracted from PPDB [74] and WordNet [75], we follow the original study [73].

The resulting output embeddings $\mathbf{W}'$ are of the same dimensionality as the input embeddings, i.e., 300-dimensional.

5.2.3 Cross–Modal Font–Word Representations

Finally, we induce a cross–modal vector space that jointly embeds both fonts and words. We start out with the font embedding matrix $\mathbf{F}$ from Section 5.2.1 and the modified word embedding matrix $\mathbf{W}'$ from Section 5.2.2.

In order to be able to connect these two spaces, we draw on the font–attribute matrix $\mathbf{M} \in \{0,1\}^{|\mathcal{F}|\times|\mathcal{A}|}$ based on the tagging data described in Section 5.1, from which we enumerate a set of pairs

$$M = \left\{ (i,j) \;\middle|\; w_j \in \mathcal{V}, i \in_{I\subset\{i|m_{ij}>0\},|I|\leq k} \sum_{i\in I} \frac{1}{\sum_{j'} m_{ij'}} \right\}. \tag{5.2}$$

Thus, for each tag w_j in our word embedding vocabulary $\mathcal{V}$, we retain the top-k fonts ranked in terms of the inverse of the number of tags those fonts have. The intuition of this filtering is that fonts with fewer non-zero entries $m_{ij'}$ in $\mathbf{M}$ tend to more specifically represent their tags compared to fonts that have numerous different tags. In our experiments, we consider different choices of k.

We construct new font and word alignment matrices $\mathbf{F}_0$, $\mathbf{W}_0$, such that the n-th row contains the font embedding from a normalized version of $\mathbf{F}$ for the font in the n-th entry in M, or the word embedding from a normalized version of $\mathbf{W}'$ for the tag mentioned in that entry, respectively. For this normalization of $\mathbf{F}$ and $\mathbf{W}'$, we first normalize each row to have a length of 1, then apply column-wise mean centering, and thereafter re-normalize each row to again have unit length [76]. To facilitate a mapping between the font and word representations, we follow a framework originally proposed for cross-lingual alignment [77]. We apply a variant of Mahalanobis whitening by computing $\mathbf{F}_1 = \mathbf{F}_0(\mathbf{F}_0^{\mathbf{F}}{}_0)^{-\frac{1}{2}}$, $\mathbf{W}_1 = \mathbf{W}_0(\mathbf{W}_0^{\mathbf{W}}{}_0)^{-\frac{1}{2}}$ so as to decorrelate different columns, as this simplifies the cross–modal mapping.

To learn a mapping, we solve what is known as the Procrustes problem, which, following Schönemann (1966) [78], can be achieved by computing a singular value decomposition (SVD) of $\mathbf{F}_1^{\mathbf{W}}{}_1$ as $\mathbf{U\Sigma V} = \mathbf{F}_1^{\mathbf{W}}{}_1$ to obtain orthogonal projection matrices $\mathbf{U}$, $\mathbf{V}$ of the two spaces into a single target space. We apply this mapping as $\mathbf{F}_2 = \mathbf{F}_1\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$ and $\mathbf{W}_2 = \mathbf{W}_1\mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}$, where $\mathbf{\Sigma}^{\frac{1}{2}}$ is incorporated for a symmetric reweighting of the columns in both matrices according to their cross-correlation. Subsequently, we apply a coloring operation that reverses the aforementioned Mahalanobis whitening, by computing $\mathbf{F}_3 = \mathbf{F}_2\mathbf{U}(\mathbf{F}_0^{\mathbf{F}}{}_0)^{\frac{1}{2}}\mathbf{U}$ and $\mathbf{W}_3 = \mathbf{W}_2\mathbf{V}(\mathbf{W}_0^{\mathbf{W}}{}_0)^{\frac{1}{2}}\mathbf{V}$.

The final cross–modal output embedding matrix $\mathbf{E}$ provides vectors for fonts $f \in \mathcal{F}$ taken from $\mathbf{F}_3$ in its first $|\mathcal{F}|$ rows and subsequently provides vectors for words $w \in \mathcal{V}$ taken from $\mathbf{W}_3$.

5.3 Zero-Shot Attribute-Based Retrieval

In this section, we describe and evaluate how our cross–modal embeddings enable zero-shot support for novel attributes. The goal is to be able to retrieve suitable fonts for a new attribute a that does not at all occur in the font–tag dataset used to induce the embeddings. In light of the incompleteness of social tags, this is an important task for the font domain. Additionally, it is also important as an indicator of the potential of our proposed multimodal discovery method, for which it serves as a key building block.

5.3.1 Method

We first obtain a cross–modal embedding matrix $\mathbf{E}$ following the three steps of our technique as described in Section 5.2. To predict the fonts associated with an attribute $a \in \mathcal{V}$, even if $a \notin \mathcal{A}$, we can consult $\mathbf{E}$ to obtain the cross–modal embedding $\mathbf{e}_a$ for a as well as the cross–modal embedding vectors $\mathbf{e}_f$ for fonts f, and simply select those fonts $f \in \mathcal{F}$ that maximize

$$\frac{\mathbf{e}_{f a}^{\mathbf{e}}}{||\mathbf{e}_f||_2 \, ||\mathbf{e}_a||_2}, \tag{5.3}$$

i.e., the ones most similar to a in terms of cosine similarity.

5.3.2 Evaluation

To evaluate this, we apply the above method for the 100 most frequent attributes a in $\mathcal{A}$ using leave-one-out cross-validation. Thus, for each target attribute a, we separately induce a different cross–modal embedding matrix $\mathbf{E}$ based only on the data for $\mathcal{A} \setminus a$, i.e., excluding a completely from $\mathbf{M}$. The above method is used to retrieve suitable fonts for attribute a without it having observed any annotations of this attribute at all.

Figure 5.8 shows top three fonts as predicted by this method for the most frequent ten attributes. As an example, for the attribute *handwritten*, representations are induced on the data excluding any tagging of fonts with the tag *handwritten*. The three fonts presented for *handwritten* are fonts with font vectors of the highest cosine similarity to our vector representation of the word *handwritten*.

The check marks next to the fonts indicate that the font is tagged with the corresponding attribute in the Web dataset, and hence the prediction is deemed accurate. The second font for *handwritten* has this symbol, confirming its accuracy. Nonetheless, as the Web dataset is known to have missing tag annotations, the lack of an association in the dataset does not necessarily mean that the prediction is inaccurate. In the case of *handwritten*, all of the three predicted fonts appear to represent the attribute, thus being accurate predictions.

To quantitatively evaluate the results in this setting, *precision* and *recall* are not well-suited, due to the incomplete tag annotations. Instead, in Table 5.1 we report

	Top 10	Top 50	Top 100
Unconstrained representations	0.31	0.33	0.22
Full method – $k = 1$ Filtering	0.46	0.30	0.19
Full method – $k = 10$ Filtering	0.46	0.35	0.23
Full method – $k = 50$ Filtering	**0.54**	**0.46**	**0.33**
Full method – $k = 100$ Filtering	0.45	0.40	0.28
Full method – $k = \infty$ (No Filtering)	0.49	0.34	0.23

Table 5.1: Mean reciprocal rank results for the 10, 50, and 100 most frequent attributes. Unconstrained representations: cross–modal embeddings based on the original unconstrained word vectors. Full method: cross–modal embeddings obtained based on modified word vectors, connected to fonts using different font filtering thresholds k.

Figure 5.7: Top three fonts for the attributes *narrow* and *wide* as predicted by cross–modal embeddings based on unconstrained original word vectors (without our modification, $k = \infty$).

the *mean reciprocal rank* for the top 10, top 50, and top 100 most frequent attributes, which is based on the rank of the first predicted font for an attribute that is also tagged as such in the Web dataset. This gives us a lower bound on the performance of our method. In addition to the results for our full method, for comparison, we also evaluate a variant of our method that omits the constraint procedure from Section 5.2.2 and instead projects regular word2vec embeddings into a common space with font embeddings, without filtering ($k = \infty$).

5.3.3 Results

Based on the results, our method retrieves fonts that are tagged with the corresponding tag in the Web dataset in very early positions of the ranked list; i.e., approximately the 2nd result for the top 50 attributes, and the 3rd result for the top 100 most frequent attributes when using top $k = 50$ fonts for training (Section 5.2.3).

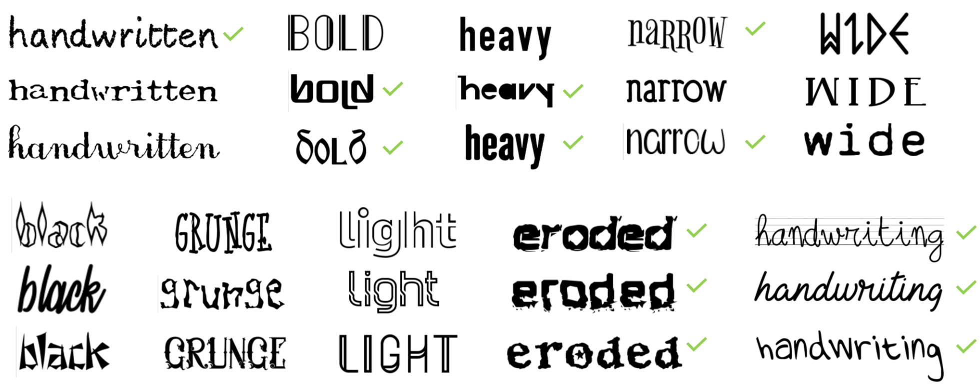


Figure 5.8: Top three fonts for the most frequent ten attributes as predicted by our zero-shot attribute-based retrieval method. The check marks represent results that are also tagged with the corresponding attribute in the Web dataset.

Our full cross–modal induction procedure results in better performance compared to the unconstrained variant, as the latter is more likely to conflate attributes with different meanings. Figure 5.7 shows the top three predictions for the attributes *wide* and *narrow* using the unconstrained variant. The fonts seem to represent attributes that are antonyms of the intended attributes. This explains the difference in performance between the two approaches.

In Figure 5.8, we observe that, in certain instances, the intended meaning of an attribute is different in the Web dataset compared to the word vectors. For example, the word *black* as a font tag is typically used to represent very thick typefaces, while based on the word vectors, it appears to be interpreted as a *dark* and *pessimist* concept in zero-shot attribute-based prediction. Note that this issue can easily be avoided if we move away from the zero-shot setting and instead incorporate a few instances of the tag into our training.

Another interesting observation is that ambiguity may affect the results in some cases. The tag *light*, for instance, is commonly used to characterize fonts with thin lines. However, in our case, the top three most similar fonts show other kinds of characteristics that may creatively exemplify being *light* in the sense of not being heavy, or perhaps are associated with *light* in the sense of lighting. Technically, distributional word vectors encode a linear superposition of all observed senses of a word [79]. Similarly, tags in social tagging platforms are often used in ambiguous ways [80].

5.4 Zero-Shot Font Similarity

We proceed to show how our cross–modal representations enable the prediction of font similarity scores in a zero-shot setting, i.e., for fonts for which we do not possess any social tag annotations. This as well is a useful building block for many font-related tasks, including our proposed multimodal discovery method.

For evaluation, we draw on a crowdsourced dataset from O'Donovan et al. [21]. In their study, in each task, a user was given a reference font and asked to select one out of two provided font options that are most similar to the reference font. The dataset (which will be referred to as $\mathcal{T}$) contains 2,340 such questions using 200 fonts, and a total of 35,287 user responses. In our experiments, we exclude questions and responses related to one single specific font for which we were not able to obtain the font file.

5.4.1 Method

We first train a cross–modal embedding matrix $\mathbf{E}$ as described in Section 5.2. For a question with a reference font f_r, and possibly similar font options f_a and f_b, we consult $\mathbf{E}$ to obtain the corresponding vectors $\mathbf{e}_{f_\mathrm{r}}$, $\mathbf{e}_{f_\mathrm{a}}$, $\mathbf{e}_{f_\mathrm{b}}$ and simply select the font

$$f =_{f\in\{f_\mathrm{a},f_\mathrm{b}\}} \frac{\mathbf{e}^{\mathbf{e}}_{f f_\mathrm{r}}}{||\mathbf{e}_f||\,||\mathbf{e}_{f_\mathrm{r}}||}. \tag{5.4}$$

	User Agreement					
	≥0.5	≥0.6	≥0.7	≥0.8	≥0.9	=1
Font	70.20	73.05	77.06	81.20	86.48	90.64
Unconstr.	70.50	73.36	77.51	81.80	87.14	90.68
Full ($k = \infty$)	70.85	73.77	77.94	82.31	87.70	91.55
Full ($k = 50$)	**70.89**	**73.86**	**78.07**	**82.44**	**88.03**	**91.78**
Oracle	81.29	85.23	89.51	93.53	97.26	100.00

Table 5.2: Individual user choice based experiment results for different user agreement levels. The column for the 0.5 agreement shows the results for the entire dataset, as the agreement cannot be lower than 0.5. Oracle shows the maximum possible accuracy, as users don't always agree.

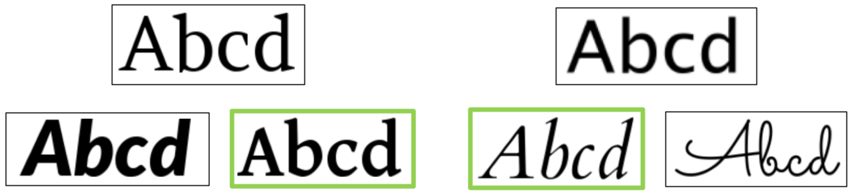


Figure 5.9: Sample high user-agreement questions that our method also agrees with (i.e., all users and our method pick the highlighted options).

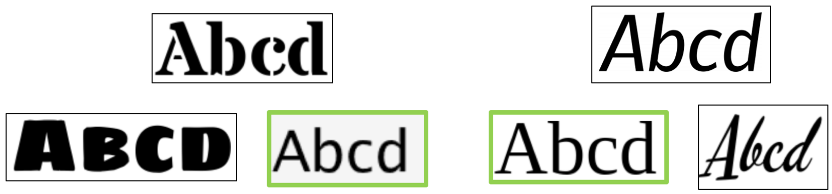


Figure 5.10: Sample high user-agreement questions that are answered differently by our method (i.e., all users select the highlighted options, whereas our method selects the others).

	User Agreement					
	$\geq$0.5	$\geq$0.6	$\geq$0.7	$\geq$0.8	$\geq$0.9	=1
Font	77.39	79.25	82.68	84.97	88.21	90.43
Unconstr.	77.78	79.45	83.10	85.67	89.08	90.61
Full ($k = \infty$)	**78.26**	80.00	83.58	86.21	89.63	91.48
Full ($k = 50$)	78.00	**80.05**	**83.71**	**86.29**	**89.96**	**91.65**

Table 5.3: Majority-choice based experiment results for different user agreement levels. The column for the 0.5 agreement shows the results for the entire dataset, as the agreement cannot be lower than 0.5. The maximum accuracy in each column is 100%, as for each question, the option with the majority of the votes is considered as the user choice.

5.4.2 Evaluation

We ensure a zero-shot evaluation setting by obtaining a new cross–modal embedding matrix $\mathbf{E}$ based on training data M that includes only tag associations for fonts $f \in \mathcal{F} \setminus \mathcal{T}$, i.e., fonts not considered in the evaluation data. For each question in the referred dataset, we can then predict the answer for the similarity question and compare it against the answers provided by the human annotators.

For some questions, users have strong agreement (e.g., all users that answered the question select the same option), whereas for others the agreement is lower (e.g., 8 users selecting option A, 7 users selecting option B). We thus analyze the performance of our method for different user agreement levels. Figures 5.9 and 5.10 show examples of questions with high user agreement, where our method agrees with the users on the examples from Figure 5.9 and disagrees with them for the ones in Figure 5.10.

We provide quantitative results in Tables 5.2 and 5.3. The results in Table 5.2 consider each user response for a question as a separate data point, and report the percentage of agreement between our method and user responses for different user agreement thresholds. In contrast, the results in Table 5.3 consider each question as a data point, and assume that the option with the majority of the user votes is deemed the correct response for that question.

5.4.3 Results

In both tables, results are compared for our original font embeddings $\mathbf{F}$ ("Font"), unconstrained ("Unconstr.") cross–modal embeddings obtained using original word2vec word vectors without the constraint-based modification from Section 5.2.2 and with $k = \infty$, and our full-fledged method to obtain the cross–modal embedding matrix $\mathbf{E}$ ("Full").

Overall, for all user agreement levels, the best results are obtained using our full method. In all but one cases, our method obtains the best results when filtering top $k = 50$ fonts as described in Section 5.2.3, compared to using all available data for training.

This shows that our method of incorporating semantic information into the visual font embeddings via cross–modal alignment yields a representation that is slightly closer to human perception.

We find that as the user agreement increases, the accuracy of our method also increases. Analyzing the disagreements, one of the insights is that users very rarely rate an all-caps font as similar to a mixed-case font, whereas our method is likely to do so, such as for the question on the left in Figure 5.10. Such preferences could be learned using a supervised font similarity method.

Our unsupervised results come fairly close to the supervised results of O'Donovan et al. [21], who were able to reach an overall individual user choice accuracy of 76.04% (where the oracle upper bound is 80.79%) on the same evaluation dataset, except for the one missing font in our experiments. Their method, however, is a supervised one that learns a similarity metric on a training fold of this evaluation dataset, and also uses a complete labeling of a set of semantic attributes of the fonts, whereas our method is completely unsupervised with regard to font similarity, and, as mentioned above, we also completely omit any available tag information about the tested fonts in order to make it a zero-shot experiment.

5.5 Multimodal Font Discovery

At this point, the results from Section 5.4 show how our cross–modal representations allow us to find similar fonts based on a reference font, while earlier, in Section 5.3, we saw how we can find fonts matching a desired attribute specification.

In this section, we show how these two notions can be combined to enable a novel form of multimodal font discovery, and demonstrate its results through sample queries.

5.5.1 Method

We create the cross–modal embedding matrix $\mathbf{E}$ as described in Section 5.2 using the Web font–tag dataset detailed in Section 5.1. For a given font $f \in \mathcal{F}$ and any suitable word $w \in \mathcal{V}$, our goal is to find fonts $f' \in \mathcal{F}$ that are similar to f and, at the same time, represent the attribute w. Note that this word w need not have occurred as a tag in our tagging dataset.

We first lookup in $\mathbf{E}$ the cross–modal representations of f and w as $\mathbf{e}_f$, $\mathbf{e}_w$, respectively, and then compute a target cross–modal representation

$$\mathbf{e}_t = \mathbf{e}_f + \mathbf{e}_w. \tag{5.5}$$

Figure 5.11: Multimodal query samples with top-1 results.

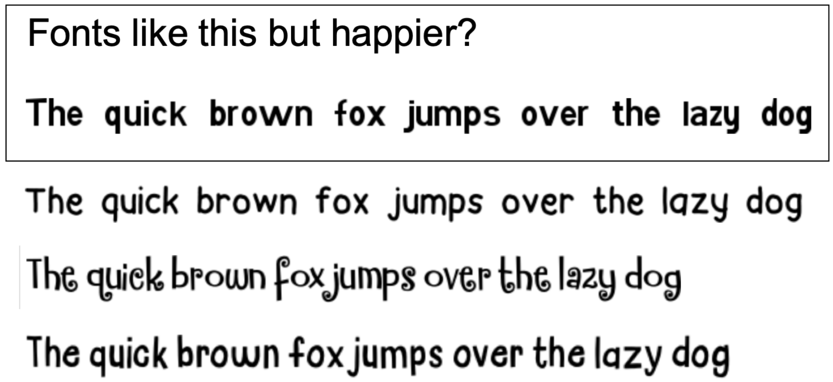


Figure 5.12: A multimodal query providing alternative directions in the top-3 results.

Given this target, we select the fonts $f' \in \mathcal{F}$ that maximize

$$\frac{\mathbf{e}_{t\,f'}^{\mathbf{e}}}{||\mathbf{e}_t||\,||\mathbf{e}_{f'}||}. \tag{5.6}$$

5.5.2 Examples

We demonstrate our method using results for sample queries. Figure 5.11 showcases sample queries that yield potentially relevant fonts as the top-1 result. The samples cover the query attributes *futuristic*, *confident*, *elegant*, *fun*, and *professional*, together with reference fonts with strong profiles. The multimodal queries are able to achieve the modifications mandated by the specified attributes while retaining the visual aesthetics of the reference fonts. In another example given in Figure 5.12, the second result appears to be significantly different from the first and third results. Yet, all results seem relevant to the query. This variety enables users to navigate in different directions in their intended search space. The examples from Figure 5.13 show that it is also possible to expand our method to include multiple attributes. This is particularly useful when the user does not have any particular reference font as a starting point but instead simply starts from a neutral default one.

Limitations. As observed in the experiments from Section 5.4, in some cases, users' perception of similarity can diverge from the embedding's notion of similarity. The top query in Figure 5.14 shows a case where an outlined reference font yields a first result that is not an outlined font. Despite the similarity between the reference font

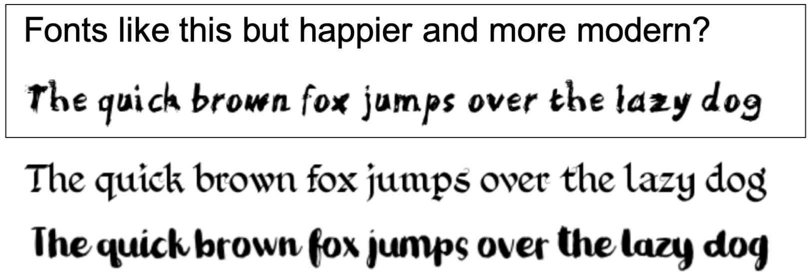


Figure 5.13: Multimodal query sample with two attributes used to modify the reference font.

Figure 5.14: Multimodal query samples with top-2 results.

and the first result, user experiments are needed to assess to what extent users would agree. Another issue is that for reference fonts that already strongly incorporate the specified attribute, the results may not seem as strong, as e.g., for the second query in Figure 5.14. Thus, further research is necessary to evaluate how users perceive the query results for reference fonts depending on how strongly the reference already reflects the specified attribute (e.g., strongly reflects, weakly reflects, does not reflect it).

5.6 Conclusion

In this chapter, we develop a cross–modal representation for fonts and words, and use it to enable zero-shot attribute-based font retrieval as well as similarity-based font retrieval. Our experiments provide insights on properties of cross–modal embeddings for fonts and words. Tag-based retrieval requires an accurate representation space that properly reflects contrasts between different attributes. Accordingly, our full method based on semantic constraints and top-k training data filtering shows improved results compared to the unconstrained baseline.

We further show that font and attribute-based retrieval can be combined by proposing a novel multimodal font searching strategy that allows the user to specify a reference font together with the changes they wish to solicit. This allows users to quickly locate new fonts that may better satisfy their design requirements.

In terms of future work, one observation, discussed in Section 5.3, is that for ambiguous words, the distribution of meanings may differ between the typographical tags and the general word embeddings. We focus mostly on semantic attributes in this study, rather than typographic ones, and see the interaction between the context of the two as a direction for future work. The work presented in this chapter was published in the proceedings of the 28$^{\text{th}}$ ACM International Conference on Multimedia [81].

Chapter 6

Affect-Aware Word Clouds

While word clouds have certain drawbacks as an analytical tool [82; 83], they are widely adopted by people of all ages and occupations for non-analytic purposes, such as preparing a gift, or introducing a topic [84]. Viegas et al. focus on a specific type of word cloud, namely those generated by Wordle, finding that a key reason for the tool's greater success in comparison with alternative word cloud tools is the "emotional impact" it creates through the tight layout it applies, and the font and color palette options it provides [84]. This is exemplified by user comments from the study that emphasize the importance of these signals, e.g.:

> "Wordles have more emotional emphasis, colors, and layouts to enhance the meaning" [84]

> "Wordles are colorful, more visually interesting, more of an emotional response and connection with the viewer than the tag clouds." [84]

For some use cases, the primary purpose of the cloud is to convey the sentiment of its input, e.g., when visualizing restaurant reviews. In such cases, the affective perception of the cloud plays a critical role. Affect, thus, is a key factor to be considered in the design of word clouds. Typography and color are two important aspects that may influence the affect evoked by a word cloud. Figure 6.1 demonstrates how different choices of fonts and colors may give rise to quite diverse affective associations.

The congruency between the textual content and the selected font has shown to have a pronounced impact on the affective interpretation of the text [9; 21]. For instance, different typefaces may result in distinct ratings for the same textual content with respect to its perceived *excitingness*. It has been observed that the response times of

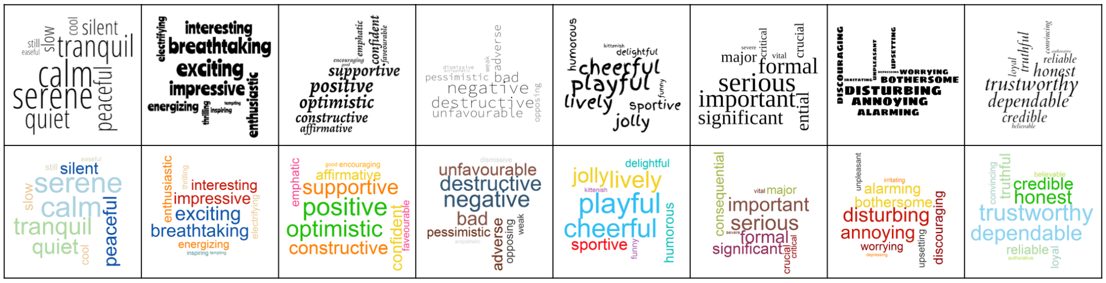


Figure 6.1: Word cloud examples using fonts (first row) and color palettes (second row) that are congruent with their message. Instead of randomly picking these paralinguistic signals as current tools do, in this study we determine their congruency with a set of eight affects, and propose a word cloud tool that helps users make congruent choices.

users decrease when fonts that are congruent with the message are used [6; 7]. Similar ties have also been observed for individual colors [85–87] and color palettes [88].

Designers routinely expend significant effort in making typographical choices that accord with the message that visual material is intended to convey. Thus, a word cloud produced by a designer would typically incorporate colors and fonts that are congruent with the message of the cloud. On the contrary, this may not be the case for non–professional users, for a variety of reasons, including a lack of such awareness, or simply an unwillingness to put in the time and effort. Despite the established relationship between typographic signals and affect, and the known impact of their congruency, word cloud tools thus far neglect to support the users in making semantically informed typographical choices. Hence the impact of automatic means of improving the congruency of fonts and colors in word cloud tools has remained under-explored. Towards the aim of enabling such support, our contributions in this study are as follows:

1. Previous studies use human annotations to solicit font–attribute ratings. We propose computational methods to obtain affective relationships based on other ratings (Section 6.2).

2. Through three user studies[1], in Sections 6.3, 6.4, and 6.5, we show that congruent fonts and color palettes, respectively, better reflect word clouds with affective content, for a range of different affects.[2]

3. We find that fonts and color palettes have complementary strengths in conveying affects, and for the majority of the considered affects, it is important to include congruent signals using both the color palette and the font.

4. Existing word cloud generators choose fonts and colors based solely on aesthetics or even randomly, neglecting their congruency with the intended affect. We instead show how they can be utilized to make typographical recommendations based on user-specified affects.

5. We further discuss a number of promising new directions for future research.

6.1 Related Work

Before describing the details of our method, we review related work on word clouds, and color palettes.

6.1.1 Word Clouds

Word clouds have been of substantial interest to the academic community, especially with regard to the employed layout algorithms [89; 90]. Several studies [91–93] focus on semantic relationships of the words, e.g., placing semantically similar words closer to each other. There are also studies aiming to create comparison clouds by combining visualizations for multiple texts [94; 95].

Wordle [96] is a widely appreciated word cloud generation tool that aims to create more pleasing word clouds ("wordles") by adopting tighter layouts, e.g., allowing a tiny word to appear within a character from a larger word, with several font and color palette options. Through a user survey and analysis of resources on the Web, Viegas

---

[1] All studies in this chapter received IRB approval.

[2] Specifically, we consider the affective attributes *calm*, *exciting*, *positive*, *negative*, *playful*, *serious*, *disturbing*, and *trustworthy*

et al. [84] analyze the users and common use cases of wordles. The results show that it attracts people of all ages, most of whom are educators or students (29%), with no other occupation accounting for more than 6%. The main reason for the tool's broad appeal is found to be the its power to create an emotional impact by means of fonts, colors, and layouts. Two other important factors are the attention-grabbing nature and the organic non-linear layout of the generated word clouds. The use cases vary from education (e.g., using Wordle to aid in introducing a topic to students) to gift-giving (e.g., creating a wordle from wedding vows), and even *guess the cloud* games[3]. Another insightful result of this analysis was that 88% of the users reported that they feel creative when using Wordle.

ManiWordle [97] enables custom manipulations in a word cloud, based on changes with regard to fonts, colors, and the layout, both at the cloud and word level. For example, one may alter the color of a selected word so as to emphasize it. Although this provides more control, users did not report feeling a strong difference in terms of creativity compared to the regular Wordle tool. WordlePlus [98] further extends ManiWordle to allow natural interaction on pen and touch-enabled tablets. EdWordle [99] facilitates multi-word editing, while preserving the neighborhood, i.e., keeping non-edited words close to their original locations. It applies a local re-wordle algorithm that re-arranges the words to close gaps.

We are not aware of detailed studies on the use of fonts or colors in text visualization from a semantic congruency perspective. A related study [100] uses font attributes, such as underlining or small caps, to distinguish set membership in set visualizations, including for emotions. Wecker et al. [101] propose using font properties such as size and color to highlight the sentiment polarity of text passages. However, they do not consider typefaces or their semantic connotations.

[3] http://guessthewordle.weebly.com/

Table 6.1: Affective attributes associated with fonts

Index	Attribute Name	Index	Attribute Name
1	calm	5	playful
2	exciting	6	serious
3	positive	7	trustworthy
4	negative	8	disturbing

6.1.2 Color Palettes

The impact of color has been studied extensively [85; 86]. Certain colors have been tied to specific forms of impact on cognitive tasks, although their effect may vary between tasks [102]. For instance, a choice of red has been found to be beneficial for detail-oriented tasks, but it does not have the same effect on creative tasks. Similar to fonts, colors have also been of special interest in marketing research [103; 104].

In contrast, the use of multiple colors together as a palette has not been studied to the same extent, as explained by Bartram et al. [88]. To address this, they conducted a study of color palette choices for a set of eight affects, which is also the source of the color palettes in our work. The study establishes several relationships between affects and perceptual color properties (hue, chroma, and lightness). As an example, the attributes *calm*, *playful*, and *positive* are found to be associated with the lightest colors. Among the core affects, *calm* is found to evoke the strongest color preferences. Other notable findings include that highly saturated light colors are not a good fit for *serious*, *trustworthy*, or *calm*, and that light colors, in general, are not very successful in conveying a *negative* sentiment.

6.2 Affective Fonts

In this section, we present our method to computationally obtain font attribute associations using crowdsourced seed data, and we evaluate that method through a user study.

6.2.1 Method

Our goal is to compute font attribute vectors $\mathbf{v}_a \in R^{|F|}$ which, for a given affective attribute $a \in A$, reflect the perception of fonts $f \in F$ with respect to the affect. The

calm	exciting	positive	negative	playful	serious	disturbing	trustworthy
calm	*exciting*	*positive*	negative	playful	serious	disturbing	*trustworthy*
calm	EXCITING	*positive*	negative	*playful*	*serious*	DISTURBING	*trustworthy*
calm	exciting	*positive*	negative	playful	serious	*disturbing*	trustworthy
calm	exciting	*positive*	NEGATIVE	*playful*	serious	disturbing	trustworthy
calm	exciting	*positive*	negative	*playful*	serious	DISTURBING	*trustworthy*

Figure 6.2: Top five congruent fonts obtained for each of the eight affects used in this chapter.

set of affective attributes A is given in Table 6.1 and was chosen because it includes the core affects in the PAD emotional state model [105] and is used in a previous study on color [88], allowing us to compare the impact of fonts and color palettes. Given the indices i from Table 6.1, we denote each affective attribute as a_i and also use the notation v_i as a shorthand for v_{a_i}, e.g., v_1 as the vector for the affective attribute *calm*. Each dimension of a particular $\mathbf{v}_a$ reflects the congruency of attribute a with respect to a different font $f \in F$.

Similar to chapters 3 and 4, in this chapter, the set F consists of 200 different fonts, taken from the aforementioned study by O'Donovan et al. [21], due to it being the largest of its sort available for download. Their data provides us with $|F|$-dimensional vectors $\mathbf{x}_t$ for a set of 31 font traits $t \in T$, providing the scores for that trait over all fonts in F. The set T of 31 traits they consider includes visual ones such as *thin*, *angular*, but also more subjective ones. We rely on the following multi-pronged procedure to obtain the desired vectors $\mathbf{v}_a$ for affective attributes $a \in A$ from this data.

Scores for "calm", "playful": Fortunately, the crowdsourced data already includes two of our eight considered affective attributes. Thus, we directly obtain $\mathbf{v}_1$ and $\mathbf{v}_5$ by selecting the relevant $\mathbf{x}_t$.

Scores for "positive": We obtain scores for the attribute *positive* by clustering all emotional traits included in the data crowdsourced by O'Donovan et al. [21]. For this, we manually filter the set of font traits $t \in T$ so as to retain only the ones of a strong emotional nature. Then we apply k-means clustering and obtain the following three clusters:

- C_1: *bad*, *boring*

- C_2: *happy*, *playful*, *attractive*

- C_3: *calm*, *charming*, *fresh*, *friendly*, *gentle*, *graceful*, *soft*, *warm*

The first observation about these clusters is that C_1 contains traits with a negative connotation, while the other two each contain positive traits. A detailed analysis reveals that C_2 includes high-arousal positive emotions, whereas C_3 includes low-arousal ones. We compute score vectors $\mathbf{x}_C$ for each positive cluster $C \in C = \{C_2, C_3\}$ as

$$\mathbf{x}_C = \sum_{t \in C} \mathrm{sim}(C, t)\, \mathbf{x}_t, \tag{6.1}$$

where the weight of each trait is calculated as

$$\mathrm{sim}(C, t) = \frac{1}{|C| - 1} \frac{\sum\limits_{\substack{t' \in C \\ t \neq t'}} d(C, \mathbf{x}_{t'})}{\sum\limits_{t' \in C} d(C, \mathbf{x}_{t'})}\,. \tag{6.2}$$

with $d(C, \mathbf{x})$ corresponding to the distance between the cluster centroid and a trait-specific vector $\mathbf{x}$.

Using these two cluster centers, the scores for the target attribute *positive* is obtained as:

$$\mathbf{v}_3 = \frac{1}{|\,C\,|} \sum_{C \in C} \mathbf{x}_C \tag{6.3}$$

Scores for "exciting", "serious", and "negative": For the attributes *exciting*, *serious*, and *negative*, we take advantage of the antonymy relationships and use scores for *calm*, *playful*, and *positive*, respectively, as:

$$\mathbf{v}_i = \mathbf{1} - \mathbf{v}_{\alpha(i)}, \tag{6.4}$$

where $\mathbf{1} \in R^{|\,F\,|}$ is a vector of ones, $i = 2, 4, 6$ are indices from Table 6.1, and $\alpha(i)$ denotes the respective antonym of affective attribute i. We follow this method based on the observation that a font that is least representative of an attribute is a candidate to best represent the opposite attribute. However, this method assumes that a font cannot be a good representative for each of the two opposing attributes, which may not always hold, given different contexts the fonts could be used in.

Scores for "trustworthy" and "disturbing": The PAD model [105] posits that complex emotions are composed of more basic ones. Following Bartram et al. [88], *trustworthy* can be defined as *positive* + *calm*. To obtain font-specific scores, we accordingly average the values for *positive* and *calm* as $\mathbf{v}_7 = \frac{1}{2}(\mathbf{v}_3 + \mathbf{v}_1)$. We performed an analogous derivation to obtain values for *disturbing* using *negative* and *exciting*, i.e., $\mathbf{v}_8 = \frac{1}{2}(\mathbf{v}_4 + \mathbf{v}_2)$.

6.2.2 Results

The top five congruent fonts for each affect in Table 6.1, as determined by the above process, are depicted in Figure 6.2. The top fonts for a particular affect appear to have similar visual characteristics, except for *exciting* and *negative*. To investigate this, Figure 6.3 plots the distribution of fonts based on the *positive* and *exciting* attribute scores. These attributes correspond to the opposites of *negative* and *calm*. For instance, the fonts that correspond to the lower end of the *positive* scale are the fonts assessed as highly *negative*. We find that the top fonts for the *exciting* and *negative* attributes have a wide range of values in the respective other scale. For example, fonts with the highest scores for *exciting* exhibit a large degree of variance with regard to their scores for the attribute *positive*. The same applies for *negative*, i.e., fonts with the lowest scores for *positive* have a wide range of scores for *exciting*. This can explain why for these two attributes, fonts with the highest congruency exhibit different visual characteristics. The top fonts for other attributes, on the other hand, are found to reside in a smaller area in the chart, i.e., have a narrower range of scores in the other scale. As an example, fonts with the highest scores for *positive* are concentrated in a region with low scores for *exciting* (between 10 and 25). This analysis is further expanded and validated in Section 6.6, where we group the top *negative* and *positive* fonts based on their scores for *calm* and *exciting*.

Overall, the majority of fonts considered are deemed highly *positive* and *calm*. Figure 6.3 also provides the categories to which the fonts belong. *Handwriting* typefaces are designed to give the impression of being hand-rendered. The characters of *monospace* typefaces occupy equal horizontal space. *Serif* typefaces have small lines attached to the end of the strokes in their characters, whereas *sans-serif* ones lack such attached lines. *Display* typefaces do not share typical typographic properties other than a lower

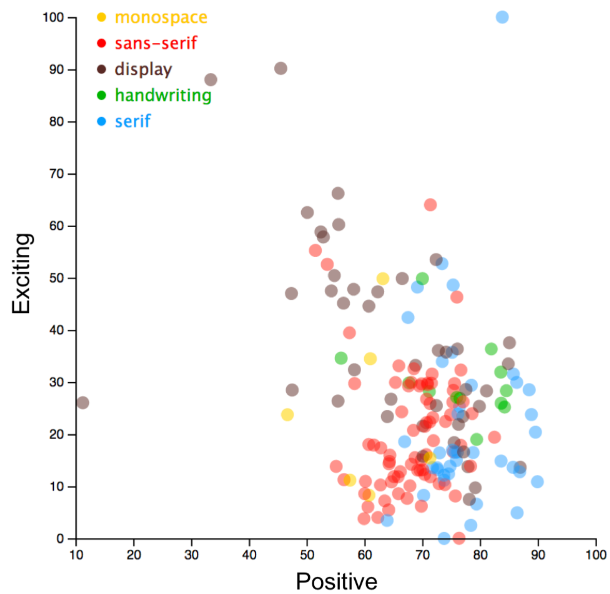


Figure 6.3: Scatter plot of the resulting font scores based on the *positive* and *exciting* attributes (scaled to $[0, 100]$ range).

Option 1	Option 2	Option 3	Option 4	Option 5
serious	serious	serious	serious	serious

Figure 6.4: A sample task from User Study I. For this sample task, the third and fifth images are generated using incongruent fonts, the second one uses a neutral font, and the first and fourth images use congruent fonts.

degree of legibility when used for body text, so they are reserved mostly for headings and other kinds of display purposes.

When we analyze the score distributions in conjunction with their font categories, we find that *Serif* fonts appear to have higher *positive* scores compared to *sans-serif* fonts. *Display* fonts are found to be *exciting*, which accords with their decorative nature. With significantly fewer instances in the dataset, *monospace* and *handwriting* fonts are scattered along a wide range of different values.

6.2.3 Evaluation (User Study I)

We evaluate the above font score computations through a user study, in which we collect user preferences among congruent, incongruent, and neutral fonts for affective word clouds.

Option 1	Option 2	Option 3	Option 4	Option 5
negative	negative	NEGATIVE	negative	negative

Figure 6.5: A sample task from User Study I. For this sample task, the third and fifth images are generated using incongruent fonts, the second one uses a neutral font, and the first and fourth images use congruent fonts.

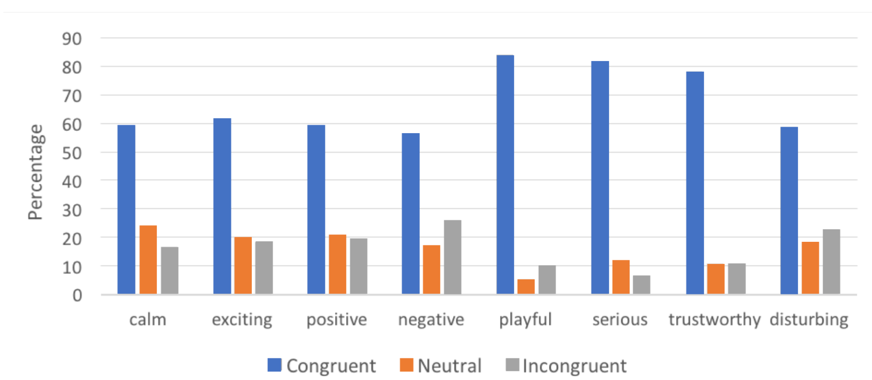


Figure 6.6: Results of User Study I, evaluating the font scores with respect to renderings of single words. The user preferences for congruent, neutral, and incongruent font choices are compared for each of the 8 affective attributes. With uniformly random selections, these, respectively, have a 40%, 20%, or 40% chance of being selected, since two out of five options are congruent or incongruent, while one option out of five uses a neutral font.

Hypothesis. Fonts with higher congruency scores for a given attribute are assessed as better representing that attribute than fonts with lower scores.

Participants and Method. We recruited 40 participants via Mechanical Turk, all from the United States, with at least 50 approved HITs and an overall approval rating of 90% or more. Participants were paid $0.01 for each task. We rely on a within-subject design, and perform counterbalancing by reversing the order of the tasks for half of the participants. The study involves 50 tasks for each participant, consisting of 6 tasks for each of the 8 affective attributes, and 2 additional validation tasks. Each task presents the name of the affective attribute using 5 different fonts, as exemplified in Figures 6.4 and 6.5. To allow for comparison with the color palette study by Bartram et al. [88], we select two congruent, two incongruent, and a neutral font. The congruent ones are selected randomly amongst the six most congruent fonts for the corresponding affect, while the two incongruent fonts are selected randomly among the least congruent six fonts. The fifth font is selected randomly among the three fonts that are in the middle of the ranked font list for each affect. The order of fonts with different congruencies are selected randomly for each task. The validation tasks include words written with 1 congruent font and 4 incongruent fonts. Participants were instructed to "*Pick the image that best represents the word*", with an additional detailed version given as "*Select the image that you think best reflects the meaning of the word shown in the images.*".

Results and Analysis. The results of our study are summarized in Figure 6.6. Across all attributes, the options with fonts determined to be congruent are frequently selected by the participants, according with our hypothesis, while the options with fonts determined to be incongruent are less frequently selected. We conducted χ^2 goodness of fit tests of user preferences for each affect based on the three font category choices (congruent, neutral, and incongruent fonts). We set significance level to 0.05. Table 6.2 provides results of these analyses, which are found to be statistically significant for each of the eight affects.

Given that the scores for *calm* and *playful* were obtained via crowd-sourcing, in our analysis, they may serve as ground truth benchmarks as to what range of scores we are to expect from high-quality human-provided ratings. Fonts rated strongly as *calm*

i	Attribute	Study I (Section 6.2.3)		Study II (Section 6.3.1)		Study III (Section 6.4.2)		Study IV (Section 6.5.1)	
		$X^2(2)$	$p <$	$X^2(2)$	$p <$	$X^2(2)$	$p <$	$X^2(2)$	p
1	calm	56.79	.001	54.14	.001	121.84	.001	48.20	< .001
2	exciting	56.33	.001	16.80	.001	44.96	.001	46.25	< .001
3	positive	48.10	.001	43.78	.001	43.13	.001	48.65	< .001
4	negative	29.03	.001	51.17	.001	89.54	.001	1.80	.4066
5	playful	195.07	.001	70.15	.001	115.02	.001	54.60	< .001
6	serious	180.03	.001	142.06	.001	43.09	.001	0.60	.7408
7	trustworthy	149.29	.001	86.68	.001	20.79	.001	14.60	< .001
8	disturbing	38.94	.001	126.69	.001	76.75	.001	27.95	< .001

Table 6.2: Chi-square goodness of fit test results for the four user studies presented in this chapter. Each affective attribute is represented by the corresponding index (i) as defined in Table 6.1. For the first three experiments, expected values are specified as 0.4, 0.2, and 0.4 for the categories *congruent*, *neutral*, and *incongruent*, respectively. For the last experiment expected values are specified as 1/3 for each of the three options. The significance level is set as 0.05 for all experiments. Results for each affective attribute in the first three user studies are found to be statistically significant, as for each analysis $p < .001$. For the last user study, all affects except *serious* and *negative* received statistically significant results. Discussions on the results can be found in the respective sections.

appear to be less preferred than those for *playful*, possibly owing to the fact that even regular fonts may also have a tendency to be perceived as calm. Indeed, the median congruency score in our data for *calm* was 76.3%, while for *playful*, it was 34.5%, confirming that neutral fonts are more likely perceived as calm.

Both *exciting* and *serious* acquire similar results to the baselines, which suggests that our method of computing their scores as reversed opposites suffices to select fonts perceived as congruent. The clustering approach used for *positive*, from which in turn scores for *negative* are derived as well, appears to yield reasonable but not overly strong ratings. While this might stem from inaccuracies in the automatic clustering, it may also be the case that it is less trivial to convey positive and negative sentiment than to convey attributes such *playful* and *exciting*. The most peculiar finding is that fonts with high scores for the attribute *trustworthy* manifest stronger preferences than those for the attributes used to compute its values (namely *positive* and *calm*). Despite being computed analogously, the ratings for *disturbing* do not exceed those for *negative* and *exciting*.

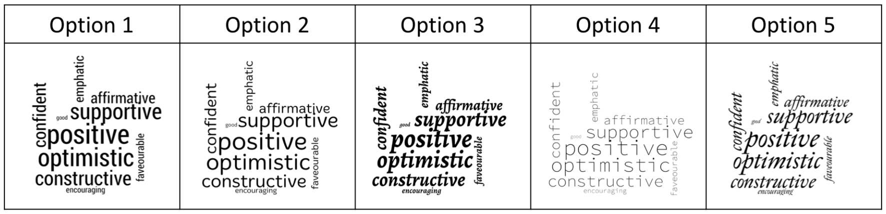


Figure 6.7: A sample task from User Study II for the affect *positive*. The third and fifth images use congruent fonts, the first uses a neutral font, and the second and fourth ones use incongruent fonts. The participants are asked to pick the image that best represents the words in the word cloud.

6.3 Affective Fonts in Word Clouds

Using the font scores obtained in the preceding section, we seek to understand the impact of the affective nature of fonts on affective word clouds. In particular, we shall determine to what extent users may prefer fonts that accord with the content of a word cloud with regard to affect.

6.3.1 User Study II

Through a user study on Mechanical Turk, we evaluate the impact of fonts on word clouds with affective content.

Hypothesis. Word clouds using fonts determined to be congruent are assessed as being more representative of pertinent affect-evoking words than word clouds using fonts determined as neutral or incongruent.

Participants and Method. The details of the participants and method of evaluation are as in Study I (Section 6.2.3), except that the displayed renderings here include words clouds instead of single words. Sample tasks are given in Figures 6.7 and 6.12. For each affective attribute, we created word clouds of 10 words, one of which is the affective attribute name itself, coupled with 9 further semantically related words to avoid confounding effects potentially caused by irrelevant words. For the same reason as earlier, each task includes 5 word clouds with the same content and layout, just using a different font. As earlier, we randomly select two congruent fonts, two incongruent ones, and a neutral one, following the procedure for Study I. Figure 6.1 provides samples of word clouds with congruent fonts from this user study.

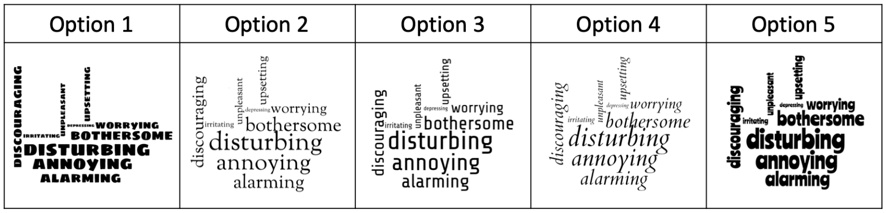


Figure 6.8: A sample task from User Study II for the affect *disturbing*. The first and fifth images use congruent fonts, the third uses a neutral font, and the second and fourth ones use incongruent fonts. The participants are asked to pick the image that best represents the words in the word cloud.

6.3.2 Results and Analysis

The frequencies of participant responses are visually presented in Figure 6.9, and chi-square goodness of fit test results are given in Table 6.2. Similar to Study I, results for each of the eight affective attributes are found to be statistically significant.

The results are in general consistent with those from the first study, as both show strong support for congruency with *serious* and *trustworthy*. Although the differences are less pronounced than earlier, across all attributes, the options with congruent fonts were chosen notably more frequently than incongruent ones, and the options with incongruent fonts were chosen less frequently than chance would predict, i.e., $\frac{2}{5} = 40\%$. It is observed that the congruent font options are more frequently selected for complex affects compared to core affects.

For *calm* and *positive*, neutral fonts were also chosen in many cases. As explained for Study I, and shown in Figure 6.3, larger numbers of fonts might appear somewhat calm or positive, and hence fonts in the middle of the ranked lists, which we assumed as *neutral*, may be more congruent.

Interestingly, *disturbing* received higher scores in this experiment compared to Study I. This may result from the layout of the word cloud, which mixes horizontal and vertical orientations, different font sizes, and different alignments, and, hence, in itself may already embody an appearance congruent with the notion of being *disturbing*. This circumstance may also explain the comparably lower scores for congruence with *calm* in comparison with Study I. We conclude that, in addition to the layout, font congruency as well merits significant consideration when designing word clouds.

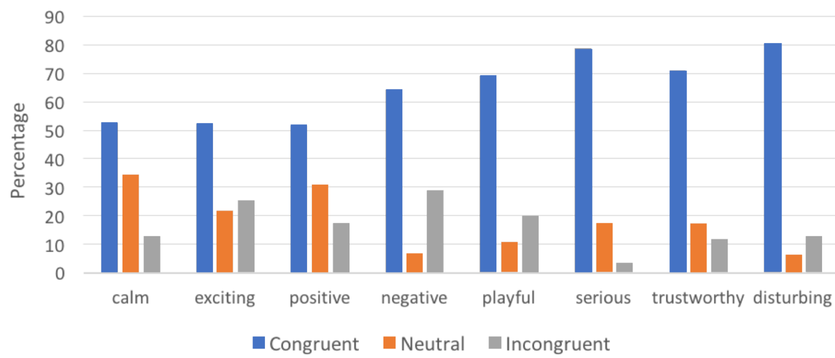


Figure 6.9: Results of User Study II, providing percentages of user responses for congruent, incongruent, and neutral font choices in the word clouds. With uniformly random selections, the expected values for the congruent and incongruent options are 40%, while for neutral it is 20%.

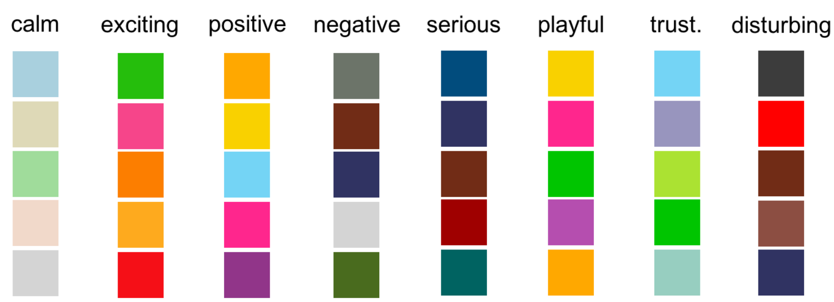


Figure 6.10: Congruent color palette samples from Bartram et al. [88].

6.4 Affective Color Palettes in Word Clouds

Apart from the choice of font, we further wish to understand the impact of affective color palettes on affective word clouds, specifically, to what extent users prefer color palettes that match the content of the word cloud affectively. To achieve this, we rely on the data from Bartram et al. [88] and carry out a user study using these palettes to create affective word clouds.

6.4.1 Data

The Bartram et al. study considers the same affect categories as this study, and their methodology to obtain the color palettes is as follows. First, 8,608 images tagged with one of the eight affective categories are analyzed to find the most common colors for

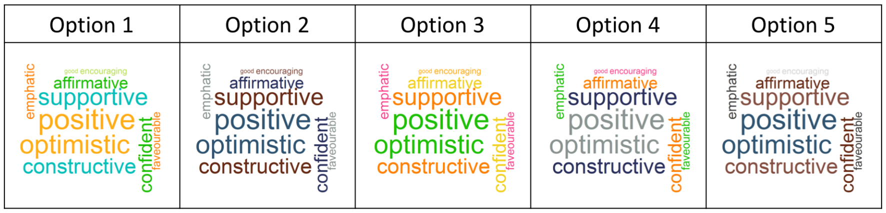


Figure 6.11: A sample task from User Study III for the affect *positive*. The second and fifth images are generated using incongruent color palettes, the fourth one uses a mixed color palette, and the first and the third images use congruent color palettes. The users are asked to pick the image that best represents the words in the word cloud.

each category. Then, with the support from a visualization color expert, a set of 41 colors is determined, which combines the most representative colors for all eight affects.

Using this set of colors, a user study is carried out requesting participants to design color palettes for one of the affect categories to be used in either a bar chart or a map visualization. The frequencies from this user study reveal the preferred colors for each affect, which are then used to create weights to come up with a palette weight concept. A weight for a palette is determined using the frequencies of the colors used in the palette. Finally, another user study is carried out to verify the results of the previous one, by generating palettes of different weights, and asking users to pick the best one. The results suggest that these weights indeed can be a good predictor of user preferences towards the color palettes.

We use color palettes from the second user study in their paper, which reveals the most preferred colors for each affect. Specifically, we rely on the palettes in Figure 7 from the paper [88] to determine the colors. Figure 6.10 a congruent color palette sample for each of the eight affect categories. The color palettes for complex affects appear to follow a pattern based on their underlying core affects. As an example, the colors for *disturbing* seem to be a combination of colors for *negative* and *exciting*.

6.4.2 User Study III

Through a user study carried out on Mechanical Turk, we evaluate the impact of color palettes on word clouds with affective content.

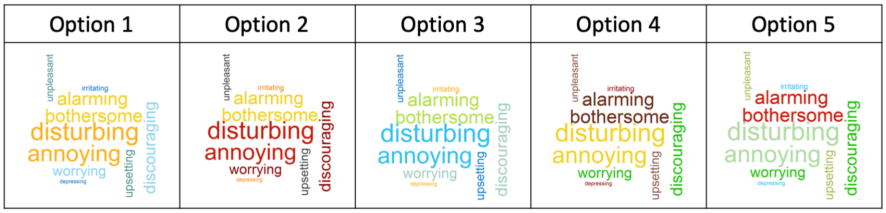


Figure 6.12: A sample task from User Study III for the affect *disturbing*. The first and third images are generated using incongruent color palettes, the fifth one uses a mixed color palette, and the second and the fourth images use congruent color palettes. The users are asked to pick the image that best represents the words in the word cloud.

Hypothesis. Word clouds using colors determined to be congruent are assessed as being more representative of pertinent affect-evoking words than word clouds using mixed or incongruent color palettes.

Participants and Method. The participant and method details are as in Study I and II (Sections 6.2.3 and 6.3.1), except that in this study, we use different color palettes in each option of the tasks, while keeping other signals, namely font and layout, the same across the presented choices. Sample tasks are shown in Figures 6.11 and 6.12. Two of the color palettes used in a task are *congruent*, two of them *incongruent*, and one is *neutral*. The procedure to create the color palettes is as follows. We use Bartram et al.'s Figure 7 [88] to obtain a list of *congruent* colors, referred to as B_i, for each attribute i as defined in Table 6.1. To generate a *congruent* color palette for attribute i, we randomly pick five colors from B_i. Figure 6.1 provides samples of word clouds with congruent color palettes from this user study. To generate an *incongruent* color palette for attribute i, we randomly pick five colors from B_j, where j is the index of the opposite attribute. We use *exciting* vs. *calm*, *negative* vs. *positive*, *serious* vs. *playful*, and *disturbing* vs. *trustworthy* as opposite pairs. The color palette for *exciting*, e.g., provides incongruent colors for *calm*. To generate a *neutral* palette, we randomly pick two colors from B_i, two other colors from B_j, and one more color from $B_i \cap B_j$.

6.4.3 Results and Analysis

As summarized in Figure 6.13, and in Table 6.2, the results are consistent with our hypothesis: Similar to the results of our previous two user studies, results for each of

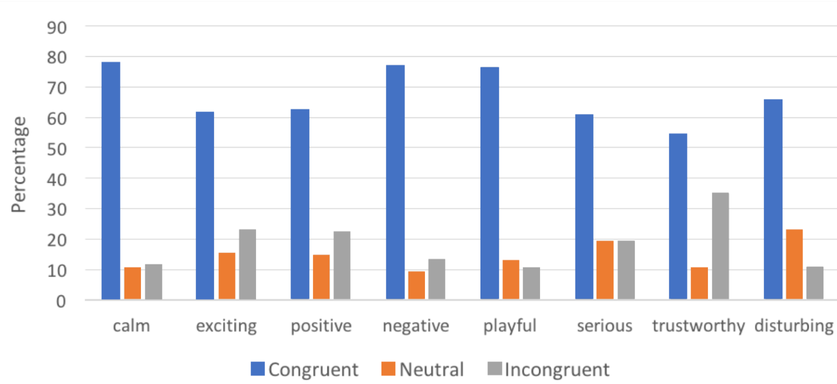


Figure 6.13: Results of User Study III, providing percentages of user preferences for congruent, incongruent, and neutral color palette choices in the word clouds. For uniformly random selections, the respective expected values for the congruent, incongruent, and neutral options are 40%, 40%, and 20%.

the eight affective attributes are found to be statistically significant. This is particularly pronounced for *calm*, *negative*, and *playful* for User Study III, while the results appear comparably less strong for *exciting*, *serious*, *positive*, and *trustworthy*. These findings mirror those from Bartram et al. [88], in which *trustworthy* and *serious* were not strongly associated with specific colors, whereas the palettes for *calm* and *playful* had highly weighted colors, reflecting a strong preference for their respective affects.

6.5 Affective Fonts and Color Palettes in Word Clouds

In Section 6.3 and Section 6.4, we carried out user studies to understand the effect of fonts and color palettes, respectively, on the affective impact of word clouds. Now we proceed to understand the relative affective power of fonts and color palettes on word clouds, and to what extent users prefer to combine these signals to create the intended affect in the word cloud. We use the same font and color datasets as described in the previous sections.

6.5.1 User Study IV

We carry out a user study on Mechanical Turk, with details similar to previous user studies. The main difference in this study is that this time we do not intend to test an hypothesis. Rather, the goal is to determine user preferences.

Participants and Method. We recruited 40 participants via Mechanical Turk, all from the United States, with at least 50 approved HITs and an overall approval rating of 95% or more. Participants were paid $0.02 for each task. We rely on a within-subject design and perform counterbalancing by creating two identical assignments with a reversed task order and invoking the reversed task order for half of the participants. The study involves 27 tasks for each participant, consisting of 3 tasks for each of the 8 affective attributes, and 3 additional validation tasks (see Figure 6.16). Each task includes 3 word clouds with the same content (10 affect-related words as described earlier) and layout.

Sample tasks are given in Figures 6.14 and 6.15. Each task includes the following three options to choose from:

- A word cloud with a congruent color palette and a neutral font (referred to as *the option with the congruent color*)
- A black-and-white word cloud with a congruent font (referred to as *the option with the congruent font*)
- A word cloud with a congruent color palette and a congruent font (referred to as *the option with the congruent color and font*)

The order of these options are determined randomly for each task. For the details of font and color palette selection, please see Section 6.3.1 and Section 6.4.2, respectively. Participants were instructed to "*Choose the image that best represents the words in the word cloud.*", with an additional detailed version given as "*Select the image that you think best reflects the meaning of the words shown in the visualization..*". From the original data, we excluded data from three participants who incorrectly answered all three validation tasks, as well as from two further participants who did not complete all tasks and incorrectly answered their validation questions. We recruited additional participants to extend the number of considered participants to 40 as planned.

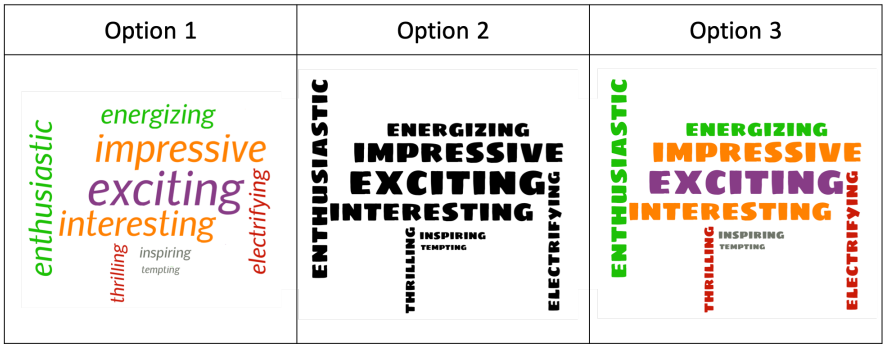


Figure 6.14: A sample task from User Study IV for the affect *exciting*. The first option shows a word cloud with a congruent color palette and a neutral font. The second options shows a black-and-white word cloud with a congruent font. The third word cloud adopts the congruent color palette from the first option, and the congruent font from the second option.

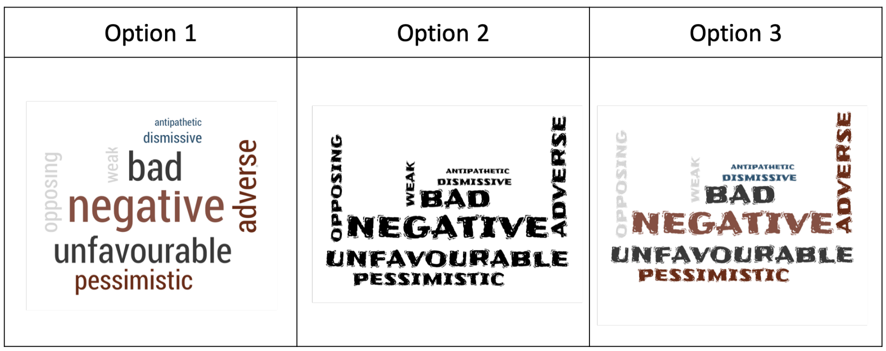


Figure 6.15: A sample task from User Study IV for the affect *negative*. The first option shows a word cloud with a congruent color palette and a neutral font. The second options shows a black-and-white word cloud with a congruent font. The third word cloud adopts the congruent color palette from the first option, and the congruent font from the second option.

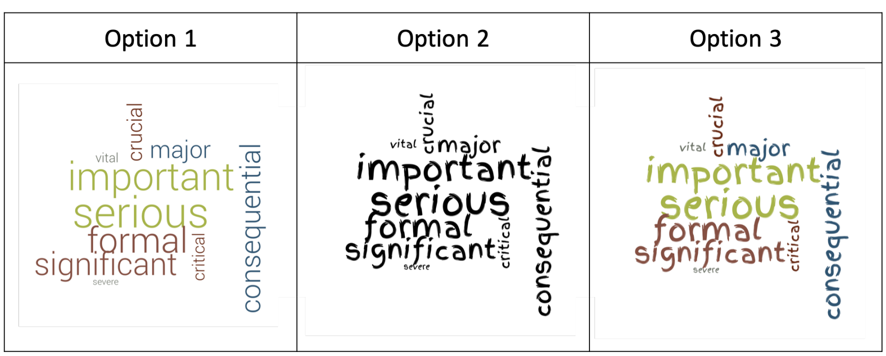


Figure 6.16: A sample validation task from User Study IV. Different from the regular tasks in the experiment, this task uses an incongruent font in two of the options, leaving the option with just the congruent color palette as the only valid choice.

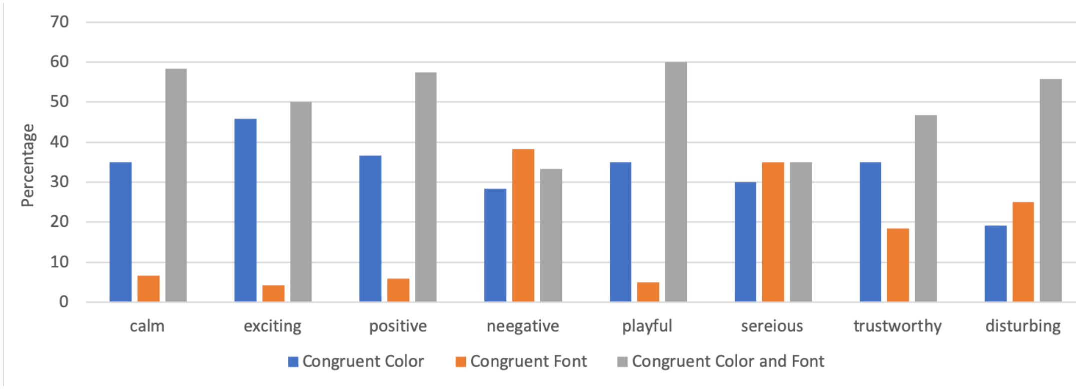


Figure 6.17: Results of User Study IV, providing percentages of user preferences for congruent color with neutral font, black-and-white with congruent font, and congruent color palette with congruent font choices in the word clouds. For uniformly random selections, the expected value for each option is $\frac{1}{3}$.

6.5.2 Results and Analysis

The results are summarized in Figure 6.17 and in Table 6.2. Based on the results of the previous studies, one may reasonably expect the option with a congruent color palette and a congruent font (*Option 3*) to receive higher preference over the other two options, in which only one of these two signals are intentionally selected as congruent with the affect. The results for *calm*, *exciting*, *positive*, *playful*, *trustworthy* and *disturbing* seem to follow this pattern. Based on the χ^2 goodness of fit analysis summarized in Table 6.2, for *negative* and *serious*, the results do not meet the significance threshold 0.05, and are thus not considered statistically significant. This may stem from the association of the neutral font with these affects (leading users to select the option with the congruent color), as can be observed in Figure 6.15. Additionally, black-and-white colors might be a good fit for these affects (leading to users selecting the option with the congruent font). For all affects except *negative*, *serious*, and *disturbing*, the option with the congruent color was more frequently selected than the option with the congruent font. This may be interpreted as the difference between a congruent color palette and black-and-white colors having a stronger bearing on the affect of a word cloud compared to the difference between the neutral font and the congruent ones.

Figure 6.18: A screenshot of our prototype Affect-Aware Word Cloud tool, visualizing an artificial placeholder text to show the affective power of fonts and colors.

6.6 Affect-Aware Word Cloud Prototype

In this section, we present our prototype tool that incorporates affects into the word cloud generation process by enabling the user to specify the intended affects such that congruent font and color palette options can be recommended. A screenshot of the tool is provided in Figure 6.18.

6.6.1 Implementation Details

User Input (Figure 6.18 Part a). The tool takes as input a text, as well as the desired affective preferences. When no preference is provided, the input is considered to be *neutral*.

In the above user studies, we used a set of words that are synonymous with the affect word to prevent any confounding impact on the evaluation. However, on real-world data, multiple affects may be evoked within a single text. Hence, one challenge in font selection is to cope with such cases.

To overcome this, we introduce restrictions on possible combinations of the affects as follows. We group the affects in two groups: core affects (*calm*, *exciting*, *positive*, *negative*) and complex affects (*playful*, *serious*, *trustworthy*, *disturbing*). Using these groups, we only allow the following selections:

1. A single complex affect.
2. A single core affect
3. Two core affects, except for opposite pairs (*positive* and *negative*, *calm* and *exciting*)

Font Selection (Figure 6.18 Part b). In this part, we describe how we handle font selections for the different affect choices listed above. When a complex affect is chosen, we use the fonts that have the highest scores for that complex affect. When just a single core affect is selected, the system as well seeks out fonts that convey that affect. However, it makes the additional assumption that since the user did not select additional affects, the user presumably does not wish to convey those other affects. As an example, when only *positive* is selected, the intended affect is presumably one that does not encompass *calm* or *exciting*. As seen in Figure 6.3, a *positive* font could be *calm* or *exciting*, so simply using fonts with high *positive* values may not be an optimal solution. Thus, when only one core affect is selected, we filter fonts so as to only retain those that exhibit neutral behavior in the unspecified affect direction. Figure 6.19 shows some examples for the affect *positive*, which includes cases where neither *calm* nor *exciting* is selected along with it, or one of them is selected. Figure 6.20 provides similar examples for *negative*. Analyzing these figures in conjunction with Figure 6.2 reveals that the fonts that are congruent with multiple affects tend to combine characteristics of fonts for each of the participating affects. We also see that while having diverse visual characteristics in Figure 6.2, the top fonts for *negative* tend to be more similar when grouped by their scores for the unspecified affect dimension, as shown in Figure 6.20. The same is also true for *exciting*.

The final scenario is that of two core affects being selected. In such cases, for each font, we ensure that the values for each of the selected affects is above a certain threshold (e.g., 50%). If they are, then we use the average of these scores to determine

Positive	Calm & Positive	Exciting & Positive
positive	calm+positive	exciting+positive
POSITIVE	calm+positive	exciting+positive
positive	calm+positive	exciting+positive

Figure 6.19: Samples of fonts with high scores for *positive* alone, and for its combinations with other core affects *calm* and *exciting*.

Negative	Calm & Negative	Exciting & Negative
negative	calm+negative	EXCITING+NEGATIVE
negative	calm+negative	exciting+negative
negative	calm+negative	EXCITING+NEGATIVE

Figure 6.20: Samples of fonts with high scores for *negative* alone, and for its combinations with other core affects *calm* and *exciting*. The visual characteristics of the fonts are similar in a combination, whereas they are somewhat more diverse between different combinations.

the congruence of the font for this affect combination. Otherwise, we use the minimum of the affect values. This technique ensures that a high congruency score is indeed reflective of a font exhibiting characteristics of *both* selected affects. If instead one were to generally consider the average, one could easily obtain fonts that reflect only one of the two affects very strongly, while having only a weak association with the other affect.

The actual font that is recommended by the tool is selected randomly among the top candidates using the above logic. The *Select* menu in the user interface, as marked by *Part b* in Figure 6.18, provides a list of recommended fonts to the user. Upon selecting a font, one may press the *Update* button to update the visualization to use the selected font. This facilitates exploring different designs for a specific affect.

Color Selection (Figure 6.18 Part c). Similar to the font selection methodology described above, in this part we explain how we handle color selections for the different affect choices. If only one affect is chosen, irrespective of whether it is a core affect or a complex one, we simply apply the corresponding color palette. If multiple core

Figure 6.21: Word clouds visualizing two restaurant reviews. The colorful cloud on the left visualizes a five-star review using a *positive* font and color palette, whereas the two star review on the right uses *negative* ones.

affects are selected, we use the corresponding complex affect, which is based on the relationships of emotions. Specifically, we use *trustworthy* for *calm* and *positive*, *serious* for *calm* and *negative*, *disturbing* for *exciting* and *negative*, and *playful* for *exciting* and *positive*. Reviewing Figure 7 from Bartram et al. [88] reveals that the color palettes for complex emotions are indeed very close to the combination of color palettes of their underlying core affects. Users can try different colors using the palettes provided in the user interface as marked by *Part c* in Figure 6.18.

Visualization (Figure 6.18 Part d). We preprocess the input to remove punctuation etc. as well as to obtain lemmatized versions of the words. To generate the word cloud, we rely on an external library[4] that is developed using the D3 framework[5]. The library allows for specifying preferred angle options for the words. It also allows for specifying a scale with which the sizes of the words are determined based on their distributions from the input. Based on the affect choices described earlier, we generate a word cloud with the automatically recommended font and color palette choices. Users have the option of trying further recommendations, or they can simply select any preferred fonts or colors.

Cloud Comparison (Figure 6.18 Part e). Our tool aims to encourage experimentation with different affective signals, as well as with different options for a specific affect. To facilitate this process, we provide the option to save a visualization in a panel

---

[4]https://www.jasondavies.com/wordcloud/

[5]https://d3js.org/

Figure 6.22: Word clouds visualizing two songs. The cloud on the left visualizes the lyrics of the song *Happy* by *Pharrell Williams* using the *positive* option. The word cloud on the right provides a visualization for the song *I hate everything about you* by *Three Days Grace*, using *negative* and *exciting* options.

on the right side of the user interface. This option obviates the need to try to remember previous designs, or save them to disk. Instead, one can rely on the comparison panel to observe the differences. In the screenshot provided in Figure 6.18, the comparison panel includes *playful*, *serious*, and *disturbing* versions of the input, respectively, from top to bottom. The rendering in the *Visualization* panel (Part d) shows the current visualization, which in this case uses *trustworthy* signals to visualize the same content.

6.6.2 Output Samples

We showcase outputs of our prototype tool using data from several domains, while demonstrating the importance of affective visualizations for these domains.

Visualizing Restaurant Reviews. Word clouds are commonly used to visualize user-written reviews, such as of restaurants. A Google web search using keywords *"restaurant review" and "word cloud"* yields around eighteen thousand results, including academic papers [106]. Since a central goal of reviews is to convey the sentiment of a customer with respect to products or experiences, the affective perception of the word cloud is significant. Figure 6.21 portrays two word clouds generated by our tool. Based on the sentiment in the reviews, the input affects are specified as *positive* and *negative*, respectively, resulting in corresponding affective visualizations. These clouds also reflect the results of our font selection method when there is only one specified affect, i.e., the input is neither *calm* nor *exciting*.

Figure 6.23: Word clouds visualizing two movies. The colorful cloud on the left visualizes the animation movie *Smurfs: The Lost Village (2017)*, using a *playful* font and color palette. The word cloud on the right is for the movie *Scream (1996)*, and, reflecting the movie's genre (horror), it is visualized using the affects *exciting* and *negative*.

Visualizing Songs. Word clouds are frequently used to visualize songs [107]. The web query *song and "word cloud"* yields 2.1 million search results. There have also been games, such as *guess the song* ones, using word clouds as their medium. Figure 6.22 provides visualizations of two songs with different affects. The word cloud on the left seeks to evoke an uplifting feeling as the original song does, whereas the one on the right aims at a *negative* and *exciting* feeling.

For song lyrics visualization, typographic signals may be particularly important due to the additional affective signals provided by the music. The words from the lyrics sometimes cannot directly be used to elicit these type of signals. The word cloud for the song *I hate everything about you* from Figure 6.22 includes the words *love* and *hate* with a similar frequency. Hence, the sentiment of the song is not clear when solely considering the words. Fonts and color palettes, thus, may be used to evoke an appropriate set of affects in such cases. This also applies to any other form of input carrying additional affective signals beyond those in the text.

Visualizing Movies. Movies are another frequent type of word cloud content. A web search using keywords *movie and "word cloud"* yields 2.7 million results, which suggests substantial interest in such visualizations. Naturally associated with emotions, movies benefit from congruent typographic signals in their visualizations. Figure 6.23

Figure 6.24: Word cloud for the United Nations using typographic signals assessed as *trustworthy*.

shows two movie word clouds created by our tool[6]. The animation/comedy movie on the left uses *playful* options, whereas the horror movie uses *exciting* and *negative* ones. It may even be possible to guess the genres of these movies just in light of the font and color choices.

Visualizing *Trustworthy* Content. Another type of word clouds we wish to specifically analyze here are *trustworthy* ones. Word clouds are frequently used to describe the values of a corporation. Searches in web sites providing professional design products, such as ShutterStock[7], return hundreds of word clouds specifically designed to reflect core values of corporations. A common goal of these corporations is to be perceived as *trustworthy*, which can be accounted for in their visualizations. Figure 6.24 provides a sample output of our tool using text[8] related to the *United Nations*, visualized using the affective attribute *trustworthy*.

[6]The input texts are movie summaries obtained from IMDB.

[7]https://www.shutterstock.com

[8]The text is obtained from the Wikipedia page for the United Nations.

6.7 Discussion

We discuss our results, their implications for users and use cases, and other potential research directions.

Affective Strengths of Fonts and Color Palettes. Across all experiments and attributes, congruent font and color palette choices were preferred by a plurality or majority, while incongruent choices were dispreferred by a majority of responses. The findings in the first two user studies have shed light on the relationship between affective responses and fonts. For attributes such as *serious* and *trustworthy*, this relationship is found to be particularly strong. Interpreting these results together with the third user study, we observe that different signals exhibit different strengths in terms of their affective impact. Based on our experiments, color palettes prove particularly powerful for *calm*, *negative*, and *playful*. Thus, a *serious* or *trustworthy* perception appears easier to evoke with fonts, whereas a *calm* or *negative* appearance can arise from appropriate color palettes. For the remaining attributes, to achieve a more pronounced effect, a combination of both fonts and colors may be a compelling option.

Supporting Creativity. In a survey [84], 88% of users reported feeling creative when using Wordle, due to the use of font and color palette options that can be explored. 81% of users reported they were trying it for fun. Hence, providing a medium in which users can feel creative ought to be an important aim of word cloud tools. We believe that allowing users to try applying different affects, or to try different options for the same affect, would substantially increase the creative potential to be explored by them, especially if they are given the chance to explore font and color palette options that are congruent with the affect(s) they specified. Nonetheless, a user study is needed to verify this hypothesis, since the users' sense of creativity is known to be hard to predict [99].

Word Clouds for Sentiment Analysis. Our output samples show that affective-aware choices of fonts are crucial for data from several domains. The same is more generally true of word clouds for sentiment analysis. There are many online resources, presentations, and academic papers [108; 109] that make use of word clouds for sentiment analysis, showcasing affective words that are used in the text. Most of them use a different word cloud for each emotion, or they use a comparison cloud to compare their intensities. In both cases, typographic signals could prove helpful to facilitate the perception of different emotions in sentiment visualizations.

Other Paralinguistic Signals and Visualizations. We have explored the affective usage of fonts and color palettes in word clouds. However, there are further affective signals to be explored in word clouds, as well as other text visualization methods to be investigated with respect to their incorporation of paralinguistic signals. For word clouds, one other such signal is the *layout*. We suspect that this could prove powerful, especially for attributes such as *disturbing* and *playful*. Again, this needs to be verified through a user study as well. In our experiments, in contrast, we keep the layout identical between different word cloud options to reduce the potential for confounding effects.

There are also several other visualizations that could benefit from an affect-aware approach. One line of such visualizations, as mentioned earlier, are visualization tools used for sentiment analysis. However, the list is not limited to this task. A tree map visualization, for instance, already conveys affect, since it relies on color palettes, and hence should perhaps be made affect-aware. In fact, for any visualization that makes use of fonts or color palettes, one may consider enabling such choices to be made more carefully and deliberately, in light of the affective nature of these signals.

Other Semantic Connections. Our study could be considered as a starting point towards exploring the connections between the semantics of the input and word clouds, or more generally, any visualization technique using the considered signals. More specific associations exist between fonts and semantic attributes [21], and these could be drawn upon to create word clouds with an even better thematic fit to the input. An example is using fonts found to be *technical* for *technical* content. Other more specific connections also exist between color and words, such as invoking the color *red* for the word *strawberry*, or *blue* for a word cloud relating to the *Smurfs*.

Regional and Cultural Differences. Previous studies reveal that regional and cultural differences affect color choices [110; 111]. A potential research direction is to enhance the word cloud tool to account for such differences based on user demographics. This might entail investigating whether such differences exist for font choices as well, especially with regard to the affective connections. Currently, in our tool, we provide users the opportunity to change the font and color choices without any restrictions, so that arbitrary personal preferences can be accommodated.

Input Sentiment Detection. Incorporating an automated sentiment detection method [112] or an emotion detection model [113; 114] into our tool could be a useful future direction. However, as discussed earlier, typographical signals are particularly useful for word clouds whose input include additional modalities apart from text, such as music. Typographical signals help encode the affective nature of these additional input signals. In light of this, detecting the sentiment of a word cloud input using only its text may not be an optimal solution. This problem can perhaps be observed in the automated sentiment detection for the movie *Smurfs*. The genre of the movie is animation, and in general it has a *playful* nature (e.g., as easily revealed by its music). However, looking at its plot summary text in isolation may lead to inaccurate predictions, especially in the absence of further background information about the movie and animation series. Hence, for the case of affect-aware word clouds, obtaining affect preferences from users, or building a tool that accepts additional inputs based on the domain of interest (e.g., music for songs, the genre for movies, ratings for restaurant reviews) could be better solutions.

6.8 Conclusion

In this chapter, we studied the affective connections of fonts and color palettes, specifically in the domain of word clouds. We have invoked a set of techniques to obtain font congruency values for several affective attributes based on a crowdsourced seed set. The results of our studies establish that such semi-automatically acquired font scores accord with human assessments of congruence, similar to previous studies that relied on human-chosen fonts.

Our findings reveal that both fonts and color palettes are potent signals in creating affective word clouds. Moreover, their respective strengths turn out to be complementary. Our last user study revealed that these signals should be used in conjunction to yield a stronger impact. Based on this work, we can conclude that fonts may be used as an additional dimension in visualizations to intentionally encode affect, and not only designers but also developers of computational tools need to account for the possibilities afforded by font and color associations with affective attributes. The work presented in this chapter was published in proceedings of the 24^{th} International Conference on Intelligent User Interfaces [115] and in ACM Transactions on Interactive Intelligent Systems [116].

Chapter 7

Conclusion

This thesis presented computational approaches to overcome the challenges associated with learning font semantics, searching fonts, and providing typographical recommendations.

In all studies presented in this thesis, we explored methods to learn font semantics using existing font–tag associations. A common theme surfaced is the successful use of CNN-based font embeddings and distributional word embeddings to capture font semantics. Using these embeddings, we created cross-modal font-word embeddings, which enabled the use of vector arithmetic for font discovery. Emotion-theoretic and lexical relationships have also been proven as useful tools to infer additional font semantics.

In Chapter 5, we proposed a multimodal font search method made possible by the cross-modal font and word embeddings. In that method, we suggested searching for a font using the similarity of a reference font together with the desired semantic direction to be added. Although it needs to be investigated through user studies, we expect this method to accelerate font discovery process. As a future direction, the proposed multimodal search approach could also be explored to improve search on other typographical signals, such as colors or color palettes, e.g., a color (palette) like this but *warmer*.

In Chapter 4, we recommended fonts for all the words in an English lexicon through shared affective connections between the words and the fonts. Our user study indicated that this is a promising direction, but its success is sensitive to the strength of the connections between the words and the emotions. We have also shown how two applications, namely *poster design*, and *word cloud design*, can benefit from semantic font recommendations.

Current text visualization approaches do not explicitly benefit from the semantic connotations of the fonts. Our work, specifically Chapter 6, is also important with its message that word clouds could make use of font semantics to help users better reflect their messages. This takes us to the conclusion that the field of *visualization* needs to take into account the semantic impact of typographical signals, which could enable improvements over existing visualization approaches, potentially leading the way to the development of brand-new ones.

Although this thesis focused solely on semantics, aesthetics and semantics work together in design products. Thus, a natural future direction is to create models that combine semantics with aesthetics [24; 117] to enable a more advanced level of user support and automation in design tools, such as for the poster design application we explored.

From short text messages to long emails, from a coffee label to huge billboard signs, from individual words to word clouds, typography impacts how people and products are perceived in every usage of text. In this thesis research, we explored some of the possibilities that can be offered by computational approaches to support users in understanding and taking advantage of this impact, concluding that it is possible to develop tools that help people better express their messages to the world.

Bibliography

[1] Christina Alm-Arvius. *Introduction to semantics.* Studentlitteratur, Lund, 1998.

[2] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Andrea Webb Luangrath, Joann Peck, and Victor A Barger. Textual paralanguage and its implications for marketing communications. *Journal of Consumer Psychology*, 27(1):98–107, 2017.

[4] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. Dances with words. In *IJCAI*, volume 7, pages 1719–1724, 2007.

[5] Paralanguage. https://en.wikipedia.org/wiki/Paralanguage. Accessed: 2020-08-20.

[6] Clive Lewis and Peter Walker. Typographic influences on reading. *British Journal of Psychology*, 80(2):241–257, 1989.

[7] Richard L Hazlett, Kevin Larson, A Dawn Shaikh, and Barbara S Chaparo. Two studies on how a typeface congruent with content can enhance onscreen communication. *Information Design Journal*, 20(3), 2013.

[8] Samuel Juni and Julie S Gross. Emotional and persuasive perception of fonts. *Perceptual and motor skills*, 106(1):35–42, 2008.

[9] Audrey Dawn Shaikh. *Psychology of onscreen type: Investigations regarding typeface personality, appropriateness, and impact on document perception.* PhD thesis, 2007.

[10] A Dawn Shaikh, Doug Fox, and Barbara S Chaparro. The effect of typeface on the perception of email. *Usability News*, 9(1):1–7, 2007.

[11] A Dawn Shaikh. The effect of website typeface appropriateness on the perception of a company's ethos. *Usability News*, 9(2), 2007.

[12] James A Mead, Rob Richerson, and Wenjing Li. Dynamic right-slanted fonts increase the effectiveness of promotional retail advertising. *Journal of Retailing*, 2020.

[13] Alysha Fligner. *The effect of packaging typeface on product perception and evaluation*. PhD thesis, The Ohio State University, 2013.

[14] Terry L Childers and Jeffrey Jass. All dressed up with something to say: Effects of typeface semantic associations on brand perceptions and consumer memory. *Journal of Consumer Psychology*, 12(2):93–106, 2002.

[15] Thomas JL Van Rompay and Ad TH Pruyn. When visual product features speak the same language: Effects of shape-typeface congruence on brand perception and price expectations. *Journal of Product Innovation Management*, 28(4):599–610, 2011.

[16] Carlos Velasco, Alejandro Salgado-Montejo, Fernando Marmolejo-Ramos, and Charles Spence. Predictive packaging design: Tasting shapes, typefaces, names, and sounds. *Food Quality and Preference*, 34:88 – 95, 2014.

[17] Carlos Velasco, Andy T. Woods, Sarah Hyndman, and Charles Spence. The taste of typeface. *i-Perception*, 6(4), 2015.

[18] Maísa MM de Sousa, Fabiana M Carvalho, and Rosemary GFA Pereira. Do typefaces of packaging labels influence consumers' perception of specialty coffee? a preliminary study. *Journal of Sensory Studies*, page e12599, 2020.

[19] Kristian Rolschau, Qian Janice Wang, and Tobias Otterbring. Seeing sweet and choosing sour: Compensatory effects of typeface on consumers' choice behavior. *Food Quality and Preference*, page 103964, 2020.

[20] Eva R Brumberger. The rhetoric of typography: The persona of typeface and text. *Technical Communication*, 50(2):206–223, 2003.

[21] Peter O'Donovan, Janis Libeks, Aseem Agarwala, and Aaron Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Trans. Graph.*, 33(4):92:1–92:9, 2014.

[22] A Dawn Shaikh, Barbara S Chaparro, and Doug Fox. Perception of fonts: Perceived personality traits and uses. *Usability News*, 8(1):1–6, 2006.

[23] Doug Fox. Free will for typeface selection: Myth or reality? *Usability News*, 12(1), 2010.

[24] Ali Jahanian. *Quantifying aesthetics of visual design applied to automatic design*. Springer, 2016.

[25] Michael Bernard and Melissa Mills. So, what size and type of font should i use on my website. *Usability news*, 2(2):1–5, 2000.

[26] Tianlang Chen, Zhaowen Wang, Ning Xu, Hailin Jin, and Jiebo Luo. Large-scale tag-based font retrieval with generative feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9116–9125, 2019.

[27] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[28] Maximilian Bloch. 1001 fonts. http://www.1001fonts.com, 2020. Accessed: 2020-08-20.

[29] dafont.com. http://www.dafont.com, 2020. Accessed: 2020-08-20.

[30] Y Wayne Wu, Michael Gilbert, and Elizabeth Churchill. "i kept browsing and browsing, but still couldn't find the one": Salient factors and challenges in online typeface selection. In *IFIP Conference on Human-Computer Interaction*, pages 225–234. Springer, 2019.

[31] James Craig and Irene Korol Scala. *Designing with type: the essential guide to typography*. Watson-Guptill, 2012.

[32] Tova Rabinowitz. *Exploring typography.* Cengage Learning, 2015.

[33] Typography. https://en.wikipedia.org/wiki/Typography. Accessed: 2020-08-20.

[34] Serif. https://en.wikipedia.org/wiki/Serif. Accessed: 2020-08-20.

[35] Monospaced font. https://en.wikipedia.org/wiki/Monospaced_font. Accessed: 2020-08-20.

[36] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[37] Saemi Choi and Kiyoharu Aizawa. Emotype: Expressing emotions by changing typeface in mobile messenger texting. *Multimedia Tools and Applications*, 78(11):14155–14172, 2019.

[38] Bianca Grohmann, Joan L Giese, and Ian D Parkman. Using type font characteristics to communicate brand personality of new brands. *Journal of Brand Management*, 20(5):389–403, 2013.

[39] John R Doyle and Paul A Bottomley. Font appropriateness and brand choice. *Journal of business research*, 57(8):873–880, 2004.

[40] Haruka Kawaguchi and Nobutaka Suzuki. Recommending colors and fonts for cover page of epub book. In *Proceedings of the ACM Symposium on Document Engineering 2018*, DocEng '18, New York, NY, USA, 2018. Association for Computing Machinery.

[41] Saemi Choi, Shun Matsumura, and Kiyoharu Aizawa. Social font search by multimodal feature embedding. In *Proceedings of the ACM Multimedia Asia*, pages 1–7. 2019.

[42] Yong Zhu, Tieniu Tan, and Yunhong Wang. Font recognition based on global texture analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(10):1192–1200, 2001.

[43] Abdelwahab Zramdini and Rolf Ingold. Optical font recognition using typographical features. *IEEE Transactions on pattern analysis and machine intelligence*, 20(8):877–882, 1998.

[44] Zhangyang Wang, Jianchao Yang, Hailin Jin, Eli Shechtman, Aseem Agarwala, Jonathan Brandt, and Thomas S Huang. DeepFont: Identify your font from an image. In *Proc. ACM Multimedia*, pages 451–459. ACM, 2015.

[45] Chang Woo Lee, Hyun Kang, Keechul Jung, and Hang Joon Kim. Font classification using nmf. In Nicolai Petkov and Michel A. Westenberg, editors, *Computer Analysis of Images and Patterns*, pages 470–477, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[46] C. Tensmeyer, D. Saunders, and T. Martinez. Convolutional neural networks for font classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 985–990, 2017.

[47] Soon-Bum Lim, Yeon-Hee Park, Seong-Kyeong Min, et al. Font recommendation system based on user evaluation of font attributes. *Journal of Multimedia Information System*, 4(4):279–284, 2017.

[48] Saemi Choi, Shun Matsumura, and Kiyoharu Aizawa. Assist users' interactions in font search with unexpected but useful concepts generated by multimodal learning. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 235–243, 2019.

[49] Qianru Qiu, Shu Watanabe, and Kengo Omura. Affective font selection: The hybrid of japanese and latin typefaces. *International Journal of Affective Engineering*, 17(2):89–98, 2017.

[50] Ken Ishibashi and Kazunori Miyata. Edit-based font search. In *International Conference on Multimedia Modeling*, pages 550–561. Springer, 2016.

[51] Tzren-Ru Chou and Jiaxuan Lin. Chinese font recommendation based on emotional adjectives. In *Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence*, pages 155–158, 2020.

[52] Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and Thamar Solorio. Let me choose: From verbal context to font selection. In *Proceedings of ACL 2020*. Association for Computational Linguistics, 2020.

[53] Saemi Choi, Kiyoharu Aizawa, and Nicu Sebe. Fontmatcher: font image paring for harmonious digital graphic design. In *23rd International Conference on Intelligent User Interfaces*, pages 37–41, 2018.

[54] Nanxuan Zhao, Ying Cao, and Rynson WH Lau. Modeling fonts in context: Font prediction on web designs. In *Computer Graphics Forum*, volume 37, pages 385–395. Wiley Online Library, 2018.

[55] Jack Qiao. Fontjoy – generate font combinations with deep learning. http://fontjoy.com/. Accessed: 2017-07-29.

[56] Shuhui Jiang, Zhaowen Wang, Aaron Hertzmann, Hailin Jin, and Yun Fu. Visual font pairing. *IEEE Transactions on Multimedia*, 2019.

[57] Zhouhui Lian Yizhi Wang*, Yue Gao*. Attribute2font: Creating fonts you want from attributes. *ACM Trans. Graph.*, 2020.

[58] H. Q. Phan, H. Fu, and A. B. Chan. Flexyfont: Learning transferring rules for flexible typeface synthesis. *Comput. Graph. Forum*, 34(7):245–256, October 2015.

[59] Paul Upchurch, Noah Snavely, and Kavita Bala. From a to z: supervised transfer of style and content using deep neural network generators. *arXiv preprint arXiv:1603.02003*, 2016.

[60] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM transactions on graphics (TOG)*, 34(4):1–10, 2015.

[61] Tugba Kulahcioglu and Gerard de Melo. Predicting semantic signatures of fonts. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 115–122. IEEE, 2018.

[62] Saemi Choi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Typeface emotion analysis for communication on mobile messengers. AltMM '16, page 37–40, New York, NY, USA, 2016. Association for Computing Machinery.

[63] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[64] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[65] Christiane Fellbaum. Wordnet: an eletronic lexical database. cambridge, massachusetts, eua, 1998.

[66] Tugba Kulahcioglu and Gerard de Melo. Fontlex: A typographical lexicon based on affective associations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[67] Tugba Kulahcioglu and Gerard de Melo. Semantics-aware typographical choices via affective associations. *Language Resources and Evaluation*, 2020.

[68] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. Web search of fashion items with multimodal querying. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 342–350, 2018.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[70] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[71] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[72] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany, August 2016. Association for Computational Linguistics.

[73] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California, June 2016. Association for Computational Linguistics.

[74] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[75] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[76] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[77] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI Conference on Artificial Intelligence*, 2018.

[78] Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[79] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

[80] Jonathan Gemmell, Maryam Ramezani, Thomas Schimoler, Laura Christiansen, and Bamshad Mobasher. The impact of ambiguity and redundancy on tag recommendation in folksonomies. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, page 45–52, New York, NY, USA, 2009. Association for Computing Machinery.

[81] Tugba Kulacioglu and Gerard de Melo. Fonts like this but happier: A new way to discover fonts. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[82] Eric Alexander, Chih-Ching Chang, Mariana Shimabukuro, Steven Franconeri, Christopher Collins, and Michael Gleicher. Perceptual biases in font size as a data encoding. *IEEE transactions on visualization and computer graphics*, 24(8):2397–2410, 2018.

[83] Cristian Felix, Steven Franconeri, and Enrico Bertini. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE transactions on visualization and computer graphics*, 24(1):657–666, 2018.

[84] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. Participatory visualization with wordle. *IEEE transactions on visualization and computer graphics*, 15(6), 2009.

[85] TW Whitfield and TJ Whiltshire. Color psychology: a critical review. *Genetic, social, and general psychology monographs*, 1990.

[86] Faber Birren. *Color psychology and color therapy; a factual study of the influence of color on human life.* Pickle Partners Publishing, 2016.

[87] Colin M MacLeod. Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, 109(2):163, 1991.

[88] Lyn Bartram, Abhisekh Patra, and Maureen Stone. Affective color in visualization. In *Proceedings of CHI 2017*, pages 1364–1374. ACM, 2017.

[89] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X Zhou, and Huamin Qu. Context preserving dynamic word cloud visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128. IEEE, 2010.

[90] Ji Wang, Kyle D Dent, and Chris L North. Fisheye word cloud for temporal sentiment exploration. In *Proceedings of CHI 2013 Extended Abstracts*, pages 1767–1772. ACM, 2013.

[91] Jin Xu, Yubo Tao, and Hai Lin. Semantic word cloud generation based on word embeddings. In *Pacific Visualization Symposium (PacificVis), 2016 IEEE*, pages 239–243. IEEE, 2016.

[92] Johann Schrammel, Michael Leitner, and Manfred Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proceedings of CHI 2009*, pages 2037–2040. ACM, 2009.

[93] Dongying Li and Xiaolu Zhou. "leave your footprints in my words"—-a georeferenced word-cloud approach. *Environment and Planning A*, 49(3):489–492, 2017.

[94] Michael Burch, Steffen Lohmann, Fabian Beck, Nils Rodriguez, Lorenzo Di Silvestro, and Daniel Weiskopf. Radcloud: Visualizing multiple texts with merged word clouds. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 108–113. IEEE, 2014.

[95] Steffen Lohmann, Florian Heimerl, Fabian Bopp, Michael Burch, and Thomas Ertl. Concentri cloud: Word cloud visualization for multiple text documents. In *Information Visualisation (iV), 2015 19th International Conference on*, pages 114–120. IEEE, 2015.

[96] Julie Steele and Noah Iliinsky. *Beautiful visualization: looking at data through the eyes of experts*. " O'Reilly Media, Inc.", 2010.

[97] Kyle Koh, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. Maniwordle: Providing flexible control over wordle. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1190–1197, 2010.

[98] Jaemin Jo, Bongshin Lee, and Jinwook Seo. Wordleplus: expanding wordle's use through natural interaction and animation. *IEEE computer graphics and applications*, 35(6):20–28, 2015.

[99] Yunhai Wang, Xiaowei Chu, Chen Bao, Lifeng Zhu, Oliver Deussen, Baoquan Chen, and Michael Sedlmair. Edwordle: Consistency-preserving word cloud editing. *IEEE transactions on visualization and computer graphics*, 24(1):647–656, 2018.

[100] Richard Brath and Ebad Banissi. Typographic sets: Labelled set elements with font attributes. In *Proceedings of the SetVR 2016 Workshop at Diagrams*, pages 29–43, 2016.

[101] Alan J Wecker, Joel Lanir, Osnat Mokryn, Einat Minkov, and Tsvi Kuflik. Semantize: visualizing the sentiment of individual document. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 385–386. ACM, 2014.

[102] Ravi Mehta and Rui Juliet Zhu. Blue or red? exploring the effect of color on cognitive task performances. *Science*, 323(5918):1226–1229, 2009.

[103] Satyendra Singh. Impact of color on marketing. *Management decision*, 44(6):783–789, 2006.

[104] Lauren I Labrecque and George R Milne. Exciting red and competent blue: the importance of color in marketing. *Journal of the Academy of Marketing Science*, 40(5):711–727, 2012.

[105] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[106] David Kamerer. Understanding the yelp review filter: An exploratory study. *First Monday*, 19(9), 2014.

[107] Michael Burch, Tobias Fluck, Julian Freund, Thomas Walzer, Uwe Kloos, and Daniel Weiskopf. Lyrics word clouds. In *Information Visualisation (IV), 2016 20th International Conference*, pages 51–56. IEEE, 2016.

[108] Saif M Mohammad. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741, 2012.

[109] Kostiantyn Kucher and Andreas Kerren. Text visualization browser: A visual survey of text visualization techniques. *Poster Abstracts of IEEE VIS*, 2014, 2014.

[110] Johannes Itten. *The elements of color.* John Wiley & Sons, 1970.

[111] Jantawan Noiwan and Anthony F Norcio. Cultural differences on attention and perceived usability: Investigating color combinations of animated graphics. *International journal of Human-computer studies*, 64(2):103–122, 2006.

[112] Xin Dong and Gerard de Melo. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of ACL 2018*, pages 2524–2534. Association for Computational Linguistics, 2018.

[113] Shahab Raji and Gerard de Melo. What sparks joy: The AffectVec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997, New York, NY, USA, 2020. ACM.

[114] Abu Awal Md Shoeb and Gerard de Melo. Are emojis emotional? a study to understand the association between emojis and emotions. *ArXiv*, 2005.00693, 2020.

[115] Tugba Kulahcioglu and Gerard de Melo. Paralinguistic recommendations for affective word clouds. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 132–143, 2019.

[116] Tugba Kulahcioglu and Gerard de Melo. Affect-aware word clouds. *ACM Transactions on Interactive Intelligent Systems*, 2020.

[117] Peter O'Donovan. *Learning Design: Aesthetic Models for Color, Layout, and Typography*. PhD thesis, University of Toronto (Canada), 2015.