MDPI

Article

Explorative Visual Analysis of Rap Music

Christofer Meinecke [1,*] **, Ahmad Dawar Hakimi** [2] **and Stefan Jänicke** [3]

1   Image and Signal Processing Group, Institute for Computer Science, Leipzig University, 04109 Leipzig, Germany
2   Natural Language Processing Group, Institute for Computer Science, Leipzig University, 04109 Leipzig, Germany; hakimi@informatik.uni-leipzig.de
3   Department of Mathematics and Computer Science (IMADA), University of Southern Denmark, 5230 Odense, Denmark; stjaenicke@imada.sdu.dk
*   Correspondence: cmeinecke@informatik.uni-leipzig.de

Abstract: Detecting references and similarities in music lyrics can be a difficult task. Crowdsourced knowledge platforms such as Genius. can help in this process through user-annotated information about the artist and the song but fail to include visualizations to help users find similarities and structures on a higher and more abstract level. We propose a prototype to compute similarities between rap artists based on word embedding of their lyrics crawled from Genius. Furthermore, the artists and their lyrics can be analyzed using an explorative visualization system applying multiple visualization methods to support domain-specific tasks.

1. Introduction

Rap music emerged from a long history and tradition as a rhetoric of resistance [1] into a standalone music genre. As of now, the hip-hop music industry is one of the biggest music industries in the US and in Germany and many other countries; it is the most streamed music on platforms such as Spotify [2]. Rap music as a part of the hip-hop culture combines the "creative use of language and rhetorical styles and strategies" [1]. This characteristic of rap music creates similarities to literature with regard to using poetic language or referencing other artists such as the rephrasing of famous quotes or from a musical standpoint through sampling. In particular, intertextuality can enhance the enjoyment of the music through emotions such as nostalgia. An artist can reference another artist by re-using a famous quote or by writing a similar line. Detecting all these of references can be a difficult task because the listener needs a lot of knowledge of the genre and its history. User-crafted annotations from platforms such as Genius.com [3] can help in this process. As references can result in similar lines, these cases can be found by similarity searches based on word and sentence embedding, as the embedding space preserves semantic relations. The similarities found can give starting points to further search for cases beyond rephrasing such as plagiarism or can just increase the knowledge of the user about the genre and its history. Through platforms such as YouTube [4], Spotify [5], or SoundCloud [6], access to new music becomes easier, with almost no barriers. This also allows one to easily copy the lyrics or other characteristics of a song. That can be hard to detect, especially for not as well-known songs that are not written in English or in the dominant languages of the country in which the plagiarist lives. Visualizations can be applied to communicate such similarities and to further ease the process of detecting them. We combine natural language processing techniques with visualizations to communicate similarities in lyrics to domain experts and casual users that are interested in music. In particular, the domain problem of detecting similar lines can be seen as a text alignment problem [7].

We hierarchically visualize the text alignments starting with an edge in a graph as an aggregate over two artists, followed by streamlines representing the songs and showing

dependencies between them and, finally, the side-by-side inspection of two lyrics in a collation manner. For this, we apply word embedding to the lyrics from Genius.com [3], which are enriched with metadata about songs, the artists, and additional annotations about the lyrics. The data are used to compute edges between the artists with weights depending on line similarities in the lyrics. Furthermore, we extended the artist graph and the alignment visualization of Meinecke and Jänicke [8] with visualizations for exploratory multi-faceted analysis of the data and an analysis of the Genius Expertise dataset [9]. For this, we designed and applied visualizations to communicate the sentiment of the lyrics of an artist, to compare the vocabulary of different artists, to compare release dates and views of artists of the same or a different genre, and to compare the development of rap genres. The visualizations can help give a better understanding of the genre and the relations between different artists, thus supporting multiple visual text analysis [10] tasks in the digital humanities such as corpus, sentiment, and text reuse analysis by distant reading methods [11]. This approach is generalizable to the lyrics of all genres of music and different languages.

2. Related Work

2.1. Similarity of Musicians

Similarity Analysis of Musicians is one of the applied use cases in the STAR about Visualizations for Musical Data by Khulusi et al. [12]. Although the text of a song is not direct musical data, it is still connected to the music and the musician.

Similarity measurements for musicians can be divided into multiple categories [13]. These categories include collaborative filtering of user data [14], compute co-occurrences of annotated tags, words on web pages [15] or biographical information [16], and content-based methods to focus on audio or textual data from the songs themselves. In our work, we focus on the lyrics of musicians; therefore, approaches focusing on user data, biographical information, web content, and sound features are not applicable.

Work using user data to measure the similarity between musicians include the use of Amazon sale statistics [17] or Spotify listing histories [18]. Similar work was conducted by Gibney [19], Cano and Koppenberger [20], and Gleich [21] based on user data and web co-occurrences. All of these methods visualize the data through graphs, focusing either on a given artist or on the whole database, but they do not include additional visualizations to inspect a more detailed level of the data.

Similar to former work, in this study, platforms such as Genius.com are crowdsourced and include rich annotated metadata about musicians and more importantly the transcribed lyrics of the artist. These text collections can be analyzed in terms of text reuse and overall similarity. Some work compared the vocabulary of rap artists extracted from Genius.com for American [22] and German artists [23]. Another work used vocabulary to define the similarity between the artists [24], which is also in the focus of our work. Still, these work only focus on the vocabulary and do not include other facets of the underlying data.

Another way to find and visualize relations between artists would be to observe the influence of the musicians of the past on currently active musicians. For example, MuzLink [25] allows for exploring collaborative and influential relationships between musical artists via connected timelines. Other work that observed this influence through graph visualizations show the history of rock [26] or try to find artists that are prototypical to a genre [27]. Similar approaches can be of interest for rap music because of the long-existing culture of referencing and collaboration, where new upcoming artists reference previous artists or are supported by established artists. There is also work in the field of music information retrieval that focuses on lyrics to compute similarities between artists but without visualizing them [28,29].

In contrast with prior work, Jänicke et al. [16] designed a visual analytics system that supports the profiling of musicians based purely on biographical characteristics and excluding their works. Similarly, Oramas et al. [30] compute artist similarities based on biographical information and word embedding.

2.2. Song Similarity

The similarities between songs are often addressed in the music information retrieval community and can be divided into context-based methods [31] and content-based methods [32]. Content-based methods focus on the audio signal, while context-based methods can include all information that is not part of the audio signal itself, e.g., metadata or lyrics. We follow a visual text analysis process and disregard content-based methods, as this information is also not available in the Genius data.

In contrast with our approach, many music information retrieval systems focus on sound features but often combine them with the lyrics [33]. Yu et al. [34] combined textual and audio features by deep cross-modal learning to retrieve lyrics based on audio and audio based on lyrics but did not include visualization. The LyricsRadar [35] allows users to browse song lyrics while visualizing their topics in a two-dimensional vector space. Furthermore, graph-based visualizations used to tackle plagiarism detection based on sound features were designed by Ono et al. [36] and De Prisco et al. [37].

2.3. Text Alignment

Our focus lies on textual data and has similarities to work based on textual alignment and more species textual plagiarism detection and text reuse. For other text visualizations of digital humanities data, we refer to the survey of Jänicke et al. [10]. Text alignment application scenarios can be divided into three areas [7]: first, collation, which examines and records similarities and differences among variant text editions; second, the detection of text reuse, such as fragments, allusions, or paraphrases; and third, translation alignments where cross-lingual connections are focused.

Common methods to visualize text reuse patterns are Grid based [38,39], Sequence-aligned [40], or Text-oriented [41] Heat Maps. More popular are side-by-side views supported by stream graphs and aligned barcodes [42–44]. Line-level variant graphs [45,46] and tabular views [47] can help visualize similarities and differences. A detailed overview of text alignment visualizations can be found in the survey by Yousef and Jänicke [7]. From a text alignment perspective, we visualize text reuse scenarios on the song- and line-levels with collation methods, where we treat similar lyrics as textual variations [43]. For this, we apply side-by-side views and variant graphs.

3. Data

For our prototype, we used a subcorpus from the Genius Expertise Dataset. The entire dataset, which was created by Lim and Benson [9], includes 223,257 songs crawled from Genius.com in the time frame between September 2019 and January 2020.

Genius.com is a website where casual users and even artists themselves can transcribe lyrics of songs and annotate them with additional information. This information can include references to other songs or artists, an explanation of specific words or phrases, e.g., slang or wordplay, or connection to historic or current events. Genius started as "Rap Genius" in 2009 but changed its name in 2014 to include knowledge for other music genres and other types of media such as literature. Through the Genius API, data about a specific artist, song, or annotation can be extracted, including metadata about other social media platforms; relationships to other songs; and other artists involved, e.g., feature guests, producers, and more. Annotations can be added by any user, but they need to be accepted and reviewed by a moderator.

The English subcorpus we used is a set of 37,993 songs by around 2300 different artists for which the lyrics were crawled. In addition, the dataset contains further metadata, annotations, user information, and information about the artists. We crawled missing metadata of the artists afterward so that these can be used for the artist profile, which can be seen in Section 4.4. Since crowdsourced data includes troll entries, the subcorpus had to be cleaned of them, resulting in a corpus with 35,783 entries. Additionally, we crawled lyrics and metadata from Genius.com of 28,969 songs by around 600 German rap artists and groups. We processed the lyrics by first removing the different section titles such as

"Intro", "Outro", and "Bridge". Furthermore, we removed punctuation, lowercased all text, and tokenized them on a word level.

4. Methodology

4.1. Tasks and Design Rationales

We formulated seven research questions for which we applied and designed multiple visualizations. We derived these questions based on the information available in the Genius data and through discussions with people interested in rap music.

- **Q1**: Which artists have collaborated, were part of the same label or group, and are similarly based on their lyrics? (Graph Section 4.3)
- **Q2**: What are the most similar artists or songs for a specific artist? (Artist Profile Section 4.4)
- **Q3**: Which songs of two artists have similar lines or are remixes, covers, or interpolations of a song? (Side-by-Side Alignments and Variant Graphs Section 4.5)
- **Q4**: How many songs has an artist released (in a certain period of time)? Which artists belong to a genre or have composed songs in a certain genre? (Scatterplot Section 4.6)
- **Q5**: When were songs that are associated with a specific genre released? (Genre Timeline Section 4.7)
- **Q6**: What vocabulary is used in a genre, by an artist, or in a song, and how does the vocabulary differ (TagCloud and TagPie Section 4.8)
- **Q7**: How is the sentiment for a specific song? Are there artists that have on average a negative or a positive sentiment? (Sentiment Barcodes Section 4.9)

For each question, a visualization can be used and different abstract tasks are performed. We created a graph with a force-directed layout where the edges between the nodes are based on the similarity of the lyrics to identify similar artists while exploring the graph. For the song and line comparison tasks, we applied a line-level alignment approach based on side-by-side views to allow for the comparison of lyrics. We used a scatterplot to show the relationship between the release date and views for each song, and they can be compared for artists and genres. The vocabulary of a song, an artist, or a genre can be inspected with TagClouds and even compared with other songs, artists, or genres with TagPies. Furthermore, we visualized the sentiment of a song as a colored barcode and the genre tags as a boxplot-inspired timeline.

The biggest challenge in the design process was to present the low-level line similarities in a way that a user can obtain an overview of the corpus quickly. Due to the corpus size, it is not possible to give a detailed overview of all of the line similarities. Therefore, we decided to aggregate the line similarities into a single value that can be used as an edge weight and, therefore, to show the relation between artists. This allows us to bridge the gap from the line level to the song level or to artist level and to encode other information, such as the social relation, into the edge.

Following Brehmer and Munzner's task abstraction [48,49], the domain-specific tasks were to *derive* references between songs; to *identify* similar musicians and songs based on their lyrics; to *explore* a network of musicians; to *compare* the lyrics, the sentiments used, and the vocabulary of different artists and songs; and to give an overview (*summarize*) of the different facets of the dataset.

4.2. Artist Similarity

We applied fastText word vectors [50] to compute similarity values between the artists based on their lyrics and thereby to include out-of-vocabulary words, which are a common phenom for rap lyrics because of the slang, adlibs, or neologism. We chose fastText because an English model trained on Urban Dictionary [51] is available. For the German corpus, we used vectors trained on Wikipedia. An advantage of using vectors trained on Urban Dictionary is a better contextualization of word vectors for slang or adlibs. We treat each line in the lyrics as a sentence, for which a sentence vector is computed through unsupervised smooth inverse frequency [52]. Therefore, the sentence vector is a weighted average of the word vectors. The weight depends on the word frequency, the vocabulary size, and the

average sentence length of the corpus. The sentence vectors are added to a faiss [53] index structure to query the lines that are the nearest neighbors based on cosine similarity. We focused on lines instead of sentences because rap artists write their lyrics line by line and because lines are often sentences. The similarity s_{ab} of an artist a to an artist b is computed based on the cosine similarity between the target line and its nearest neighbors:

$$s_{ab} = \sum_{i=0}^{n} \sum_{r=1}^{k} cos(l_i, l_r) \cdot ((k+1) - r)$$

For two artists, we used all of their lines that are nearest neighbors, i.e., the most similar ones in the corpus. For such a pair of lines l_i and l_r, we took the cosine similarity $cos(l_i, l_r)$ multiplied by the number of nearest neighbors $k + 1$ minus the rank r of the neighbor l_r. Through this, we obtained rank-based weighting. We then took the sum over all such pairs for two artists. This value was further normalized by using the total number of lines of all songs of the artist. For two artists a and b, this results in two similarities s_{ab} and s_{ba} because the nearest neighbor relations between vectors is not a symmetric relation. The similarity values are summed together to obtain a single edge weight for the graph. Furthermore, we apply a Box–Cox Transformation [54] and a min–max normalization to the edge weights to obtain similarity values that are easier to interpret by humans and that are between 0 and 1. We apply a Box–Cox Transformation because it transforms a distribution into one that is close to a normal distribution. The resulting distribution can be seen in Figure 1c, and the original skew distribution of the edge weights can be seen in Figure 1d.
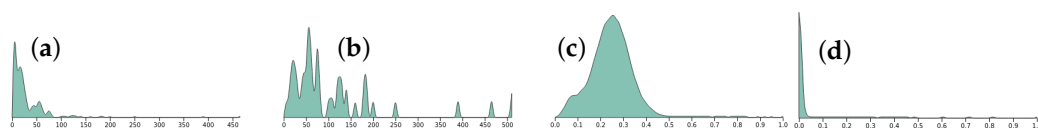

Figure 1. Kernel density estimate plots of the German graph, showing the minimum number of songs (**a**), the maximum number of songs, (**b**) the edge weights after min–max normalization with Box–Cox Transformation (**c**) and without (**d**).

4.3. Artist Similarity Graph

Following the Information Seeking Mantra [55], we started by giving an overview by visualizing the similarities between the artists as a node-link diagram. For this, we represent each artist as a node and use the artist similarity as the edge weight, so an edge indicates that two artists are similarly based on their lyrics.

Design For the graph, we chose a force-directed layout, as they are easy to understand, flexible to use with graph aesthetic criteria, and easy to interact with for a user to change the positions of the nodes. To reduce visual cluttering, the user can filter the displayed edges with sliders. For this, the similarity values and the minimum and the maximum number of songs of an artist can be used. The distribution of the edge weights and the distributions of the minimum and the maximum number of songs of the artists that are connected by the edges are displayed as kernel density estimate plots, which can be seen in Figure 1a–c. This allows a user to visually assess the impact on the displayed edges in the graph. We use a bandwidth of 1 to create a smooth estimation of the distribution.

After filtering, all nodes without an edge satisfying the conditions are removed. Furthermore, we color-coded the edges to show different relations. A blue edge indicates that two artists have at least one song together, a purple edge indicates that the artists are or were signed by the same label, an orange edge is a "part of" relation for group members, while a red edge shows an unknown relation. We chose red for the unknown relations to better highlight them as they represent an unknown or missing social relation. The different relation types show social connections beyond the lyrics, which can give hints on why the lyrics of the two artists are similar. We extracted the relation types from the lyrics and the Genius.com metadata; for some cases, we added the label or "part of" relation with

domain knowledge. Furthermore, we mapped the similarity value of two artists on the edge thickness to highlight relations with a higher similarity value.

Next to the graph, a list of the most similar song pairs is displayed. The song pairs are color-coded from white to red on a linear scale depending on the number of nearest neighbors. When clicking on a song pair in the list, the side-by-side alignment view is displayed (Section 4.5). Furthermore, a user can search for two specific artists of interest or click on an edge in the graph to investigate the songs of the artists.

Use Case A sub-graph of the German corpus can be seen in Figure 2, showing different clusters. The nodes at position (a) show previous members of the German label "Aggro Berlin" and the rap crew "Die Sekte" such as Sido, TonyD, and BTight. At position (b), previous members of "Ersguterjunge" and "Berlins Most Wanted" can be seen such as Bushido, Fler, Kay One, Eko Fresh, Nyze, M.O.030, and Baba Saad. Above these artists, more Berlin-based artists such as Prinz Pi can be seen. Another interesting thing to notice is that Eko Fresh is connected to a large number of artists, showing his influence on the German rap scene through collaborations and supporting new artists. Position (c) shows Hustensaft Jüngling and other artists that he collaborated with. Some of the edges were created because of the exhaustive use of brand names such as Gucci or Fendi, or drugs such as Lean. At position (d), the Hamburg-based group's Beginner and ASD, with some of their members, can be seen. Positions (e) and (f) show Frankfurt-based artists such as AZAD, Jonesmann and Jeyz, and Haftbefehl with his brother Capo and label member Soufian. Label members and feature partners of Capital Bra can be seen at position (g). Besides these examples, multiple "part-of" and feature relations can be found.

Another sub-graph showing the English corpus can be seen in Figure 3. Position (a) shows the $uicidebo$ and Three 6 Mafia, where $uicidebo$ reused multiple lines from Three 6 Mafia songs, which can also be seen in multiple entries in the list of most similar songs in Figure 4a. At position (b), Migos and two of their members, Offset and Quavo, can be seen together with multiple feature partners. Around Young Thug (c) are artists he collaborated with such as Lil Uzi Vert, who he influenced, and Lil Keed and Gunna, who were signed by his label "YSL Records". Position (d) shows multiple artists from Chicago that are associated with the Drill genre such as Lil Durk, Lil Reese, Chief Keef, and his cousin Fredo Santana.

4.4. Artist View

An interesting property of the Genius.com data is the rich annotated metadata including references and information about the artists. We give an overview of some of the metadata from Genius.com and display a list of the most similar artists based on their lyrics and all of the songs of the artists in the artist profile view. This view is accessed when clicking on a node in the graph or the artist's name in the side-by-side view.

Design The list of most similar artists is color-coded in the same way as the graph but instead of the edge thickness, saturation is used. Through this list, the user can further explore other artists. The list of songs includes the ten nearest neighbors color-coded in the same way as the list of the most similar songs for each song. Furthermore, the metadata from Genius.com are used to display relations with other songs. These relation types are samples, sampled in, interpolates, interpolated by, cover of, covered by, remix of, remixed by, live version of, and performed live as. By clicking on a color-coded nearest neighbor, the alignment view pops up. Therefore, a user can explore the network and find different points of interest to further investigate the alignments.

Use Case Figure 4b shows the profile of Berlin based rap crew BHZ. Similar artists show artists that are either from Berlin or have at least one song collaborated with them. Below, we can see that the song "LSD" is an interpolation of "Saphir" by Yung Kafa and Kücük Efendi.
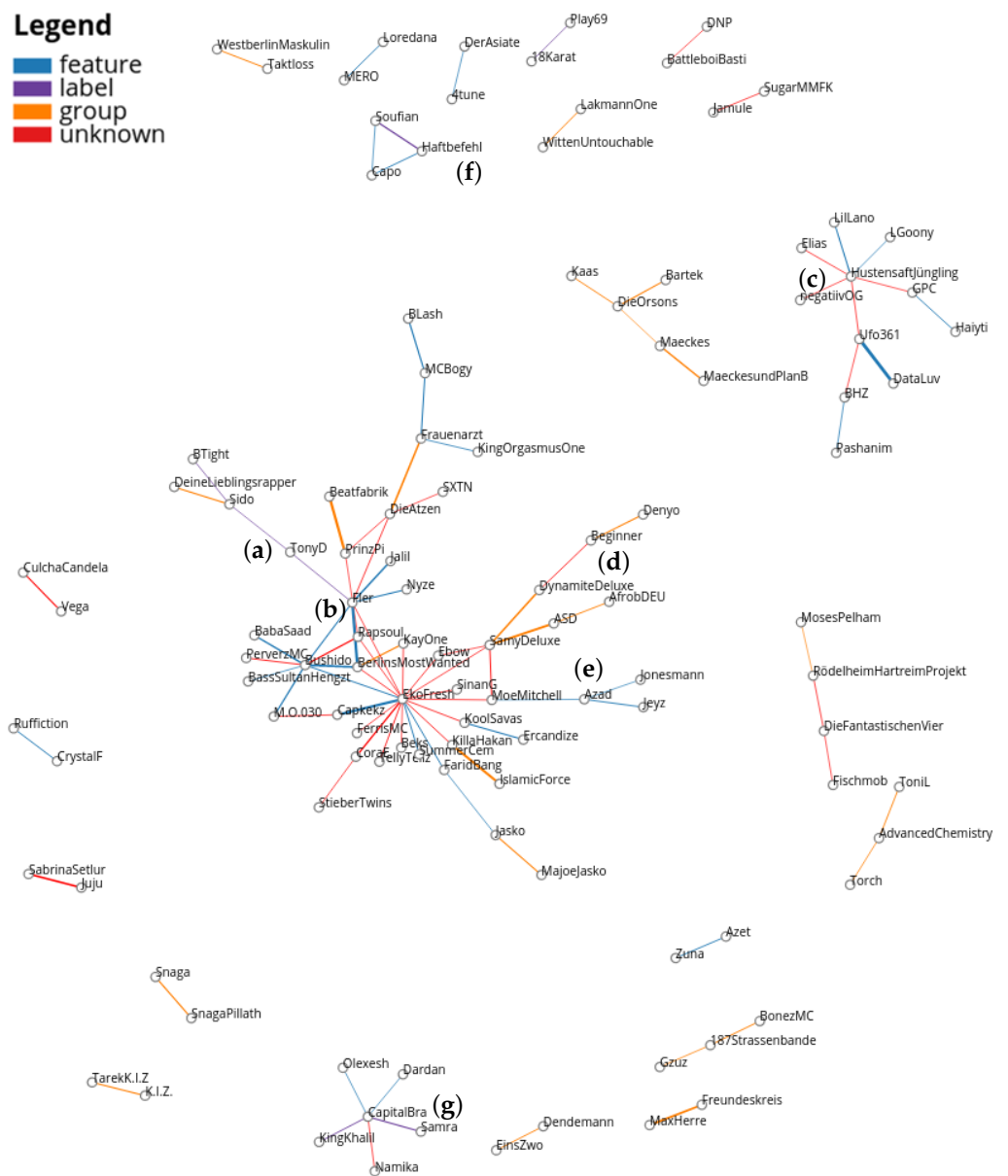
Figure 2. An excerpt of the similarity network of German rap artists based on the most similar lines in their lyrics. Label and collaboration partners tend to be connected. (**a**) previos members of the label "Aggro Berlin" and the rap crew "Die Sekte", (**b**) previous members of "Ersguterjunge" and "Berlins Most Wanted", (**c**) Hustensaft Jüngling and some feature partners, (**d**) Hamburg-based artists, (**e**,**f**) Frankfurt-based artists, (**g**) Capital Bra with label and feature partners.

4.5. Monolingual Alignments

The nearest neighbor relation can be used to compare the songs of two artists of interest on different levels, on a song level showing all relations between two artists and on a line level showing the exact nearest neighbor relations of two songs.

Design We use stream graphs to visualize the nearest neighbor relations between songs. For this, the number of nearest neighbors is mapped on the saturation of the edge between two songs. To reduce the visual cluttering, a filter mechanism can be applied. A user can filter based on the number of nearest neighbors and the release date of the songs.

Figure 3. An excerpt of the similarity network of English rap artists based on the most similar lines in their lyrics. Collaboration partners tend to be connected. (**a**) $uicidebo$ and Three 6 Mafia, (**b**) Migos and two of their members with feature partners, (**c**) artists that collaborated with Young Thug or were signed by his label, and (**d**) multiple artists from Chicago and associated with the genre Drill.

The lyrics of the songs can be read when clicking on a streamline. Both song lyrics are placed side-by-side while the nearest neighbors of each line are shown similar to the visual analytics system iteal [43]. This allows a user to read the lyrics side-by-side while investigating the alignments. Each alignment is visualized as a streamline connecting the lyrics. Furthermore, the user can filter the alignments based on a slider. The filter values correspond to the cosine similarity between the lines in the alignment. This allows to further investigate the nearest neighbors of two songs of interest. When clicking on a streamline of interest, the alignment is visualized as a variant graph using TraViz [45]. Furthermore, all of the

lines that are nearest neighbors of both lines are shown with TraViz, as seen in Figure 5. This highlights reused or shared words between the lines. These nearest neighbors can be used to move to another song pair of interest where the alignment occurred.
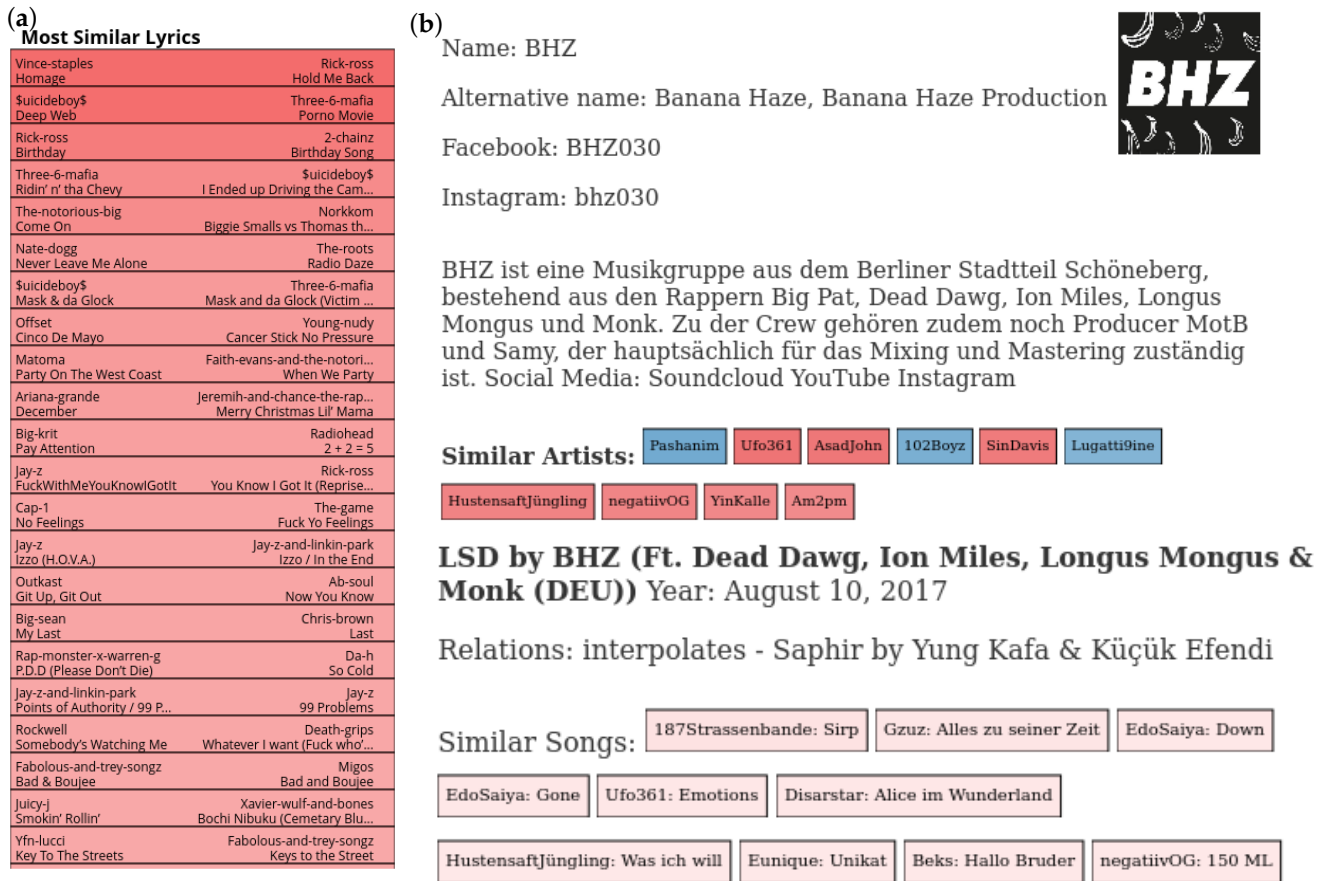

Figure 4. The most similar English songs (**a**) and the artist profile of the German rap group BHZ (**b**).

Use Case Figure 6 shows an example of the song similarities of two artists. In this case, the songs of the $uicideboy$ and Three 6 Mafia are displayed side-by-side. Some of these pairs can be found in Figure 4a and are part of a lawsuit that was filed by Three 6 Mafia against $uicideboy$ [56]. For these songs, samples, or lines that are part of the hook were reused.

Examples for the monolingual alignments of two songs can be seen in Figure 7. Figure 7a shows "Kool Savas - Komm mit mir" and "Alligatoah-Komm mit uns", where Alligatoah parodies the original song by Kool Savas. The excerpt in Figure 7b shows "Sido-Du bist Scheiße" and "Tic Tac Toe-Ich find dich scheiße", where the song by Sido sampled the original by Tic Tac Toe.

4.6. Scatterplot

To give an overview of our dataset, we used a scatterplot. It shows the release date and the number of views on the Genius.com page of the corresponding songs.

Design The visualization allows for exploring the data by zooming and panning. A mouseover over a single data point displays more information about the song. Clicking on it generates a TagCloud based on the lyrics of the song, which represents a summary of the song. The genius expertise dataset also contains genre tags about each song. Thus, it is also possible to filter by artist and by genre tags. It makes it possible to see at which time which artist was active and which genre was predominant. There is also the possibility to compare artists and genres with each other. To simplify the comparison, we use different colors and if a song has both tags, it is colored with a third color.
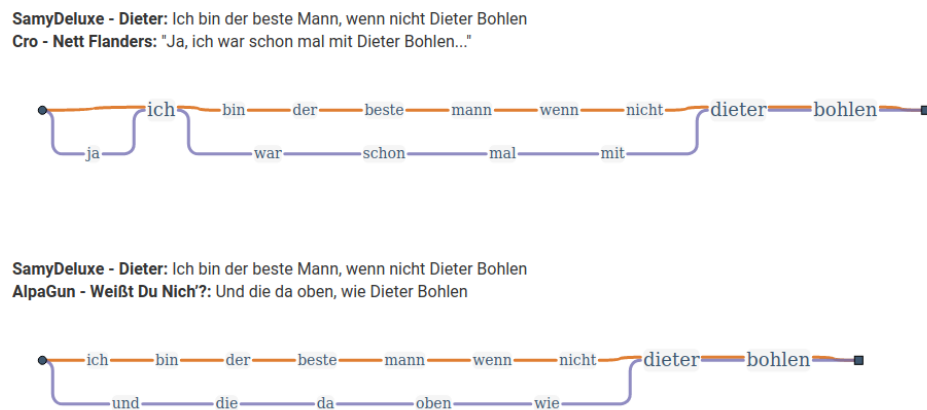
SamyDeluxe - Dieter: Ich bin der beste Mann, wenn nicht Dieter Bohlen
Cro - Nett Flanders: "Ja, ich war schon mal mit Dieter Bohlen..."

SamyDeluxe - Dieter: Ich bin der beste Mann, wenn nicht Dieter Bohlen
AlpaGun - Weißt Du Nich'?: Und die da oben, wie Dieter Bohlen

Figure 5. Two of the nearest neighbors of a line by Samy Deluxe displayed with TraViz.

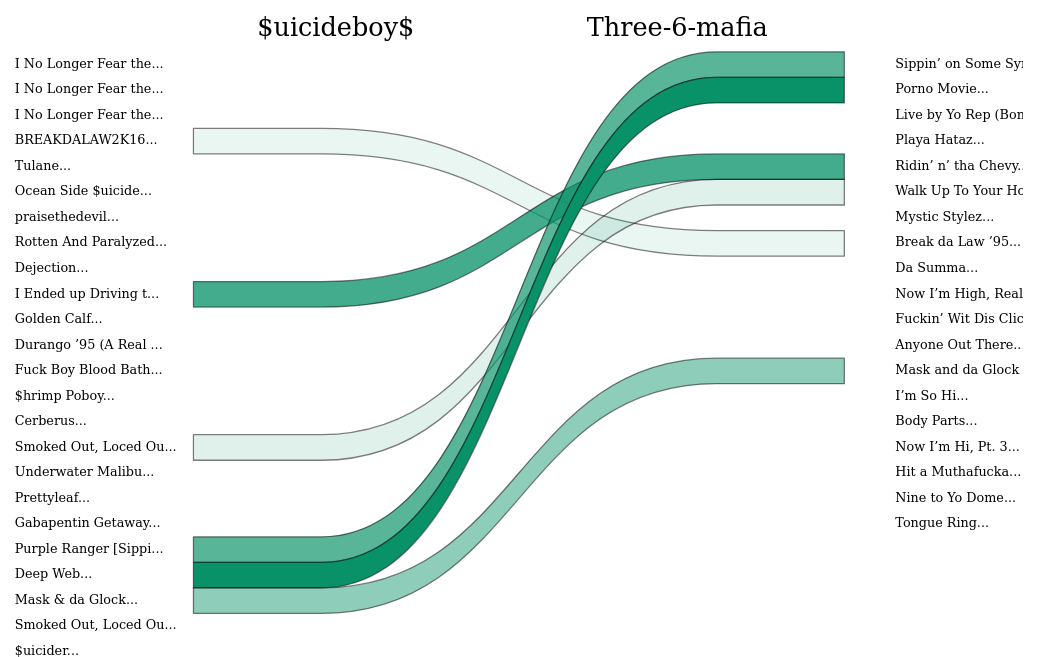**Figure 6.** Excerpt of the song level of $uicideboy$ and Three 6 Mafia. All connected songs show cases where $uicideboy$ reused lines from Three 6 Mafia.

Use Case In Figure 8, you can see the comparison of two artists. Each green dot represents a conscious hip-hop track by Kendrick Lamar and each orange dot represents a trap song by Drake. The increase in the released trap music in the last decade can also be seen in Drake's songs.

4.7. Genre Timeline

Rap music can be divided into multiple different subgenres. In order to see the development of these subgenres, we display them on a timeline based on the annotations of the Genius Expertise dataset [9]. First, we filtered the genre tags to exclude non-rap tags and non-English tags, resulting in around 40 genres.

Figure 7. Monolingual alignments on the line level of the songs "Kool Savas-Komm mit mir" and "Alligatoah-Komm mit uns" (**a**), and "Sido-Du bist Scheiße" and "Tic Tac Toe-Ich find dich scheiße" (**b**).
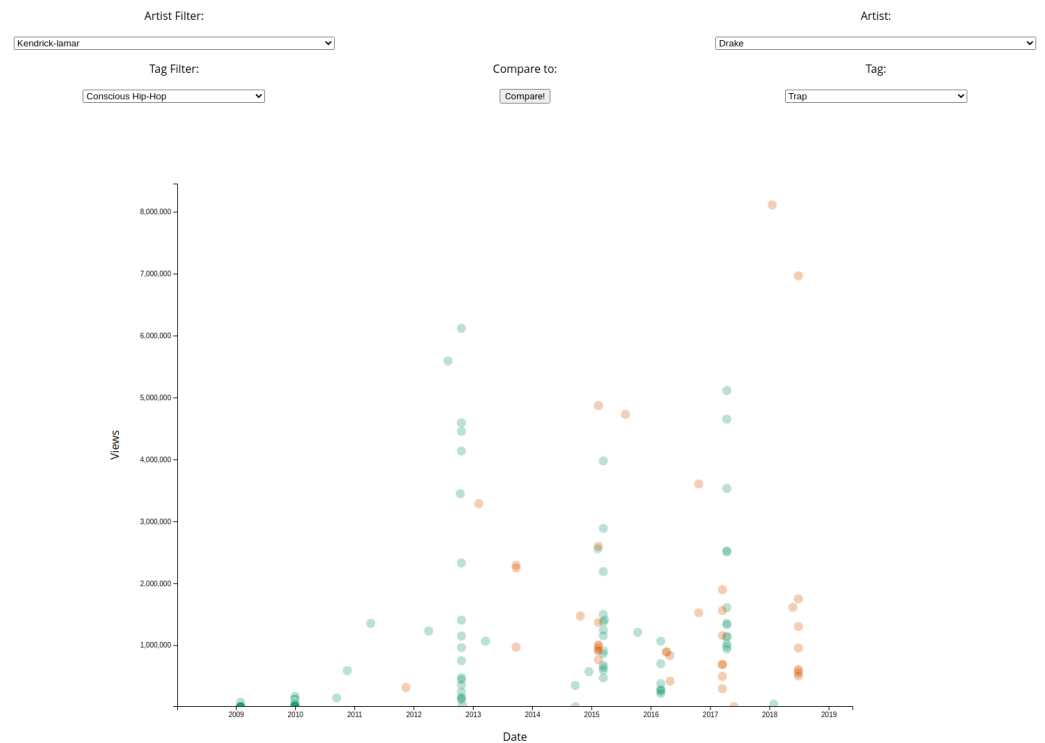

Figure 8. The scatterplot compares Kendrick Lamar's conscious hip-hop tracks with Drake's trap songs from 2009 to 2019 in terms of views and release date.

Design For each genre, we computed a boxplot representation with lower and upper whiskers, lower and upper quartiles, and the median of the release dates of the songs annotated for a genre. The oldest release date of a genre is used as the endpoint of the

lower whisker, while the newest release date is used as the endpoint of the upper whisker. The median is encoded as a colored circle. The range between the lower whisker and the lower quartile shows where the first 25 percent of release dates are located, and the range between the lower quartile and median show the next 25 percent. The same goes for the range between the median and upper quartile, and the upper whisker and upper quartile. In the visualization, we sorted the genre tags by the earliest release date of a song. We used colors to better differentiate between the genres, but they do not convey a similarity between genres.

Use Case For example, we can see in Figure 9 that, for the genre Dipset, the lower whisker and the lower quartile and the upper whisker and the upper quartile are the same. Even more, the distance between the lower quartile and the median (2004–2006) is smaller than the distance between the median and the third upper quartile (2006–2016), which shows that 50 percent of the songs were released in a short time period between 2004 and 2006. The Dipset movement goes back to The Diplomates, a hip-hop crew that released their first studio album in 2003, which influenced a lot of international artists in the following years. Furthermore, for many genres, the distance between the lower whisker and the lower quartile is significantly larger than the distance between the upper quartile and the upper whisker. This shows that there are more songs in the dataset that are newer than there are older songs, which can hint that either there are more missing older songs than newer songs and/or that more rap songs are published over the last years.
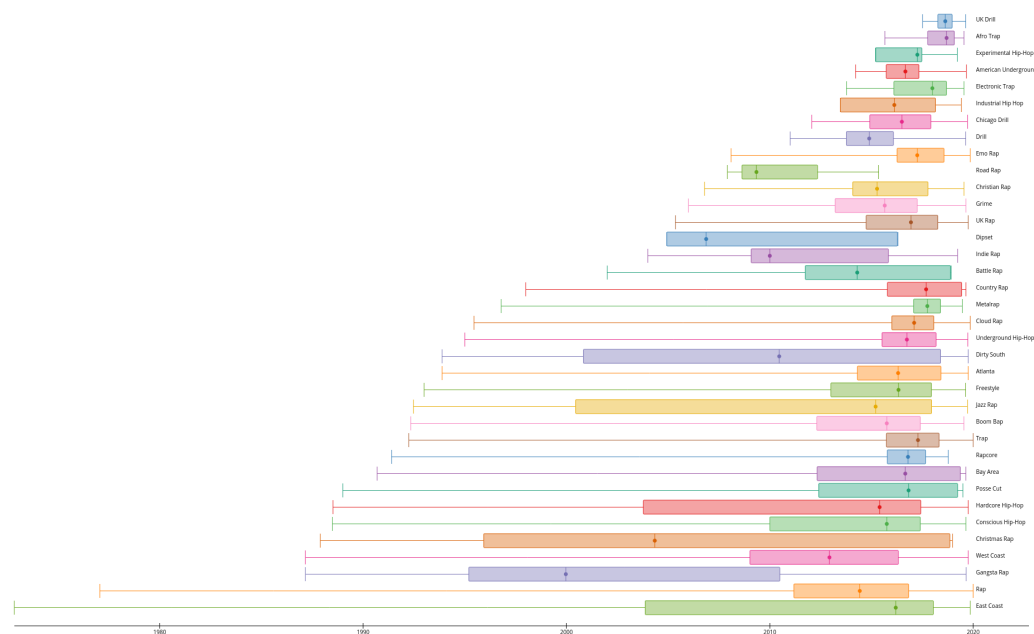

Figure 9. A timeline visualizing all rap genres associated in the Genius Expertise dataset. Each genre is shown as a boxplot.

4.8. Compare Vocabulary

In order to compare the vocabulary of the artists, we allow a user to select multiple artists to visualize the vocabulary in a TagPie [57] before the visualization stopwords are removed from the vocabulary and all words are lemmatized.

Design The word font size is mapped on a logarithmic scale so that the smallest value is mapped to 10 and the largest value is mapped to 50. Furthermore, we allow for multiple select options. The user can change the number of tags that are shown, the applied TagPie style, and the measurement that is used for the font size in the visualization. The measurement for a word w and an artist a can be either $f_w(a)$, $y_w(a)$ or the z-score, i.e., $z_w(a)$

$$y_w(a) = f_w(a) - \sum_{a_i \in A} f_w(a_i)$$

$$z_w(a) = \frac{f_w(a) - \mu_w}{\sigma_w}$$

$f_w(a)$ is the number of times that a word w is used by an artist a. $y_w(a)$ is defined by subtracting the number of occurrences over all artists $a_i \in A$ from $f_w(a)$, and the z-score denotes that the number of standard deviations $f_w(a)$ is below or above the mean value for the word w in the whole corpus. While $f_w(a)$ only shows the number of occurrences, $y_w(a)$ can be used to highlight words that are unique to an artist in the corpus or rarely used by other artists. Similarly, the z-score allows for detecting words that are common for a group of artists but more rarely used by others. An example for $f_w(a)$ and the z-score can be seen in Figure 10.

Use Case Taking only high-frequency words without normalization as in Figure 10a, we see that we obtain a lot of generic words that are often used in old-school hip-hop tracks. These do not show the wordiness of the individual artists of the Wu-Tang Clan, using the z-score results to not obtain the generic words and instead more specific frequently used words. For example cream (C.R.E.A.M.-Cash Rules Everything Around Me) is the most streamed and best-recognized song of the group. Starks stands for Tony Stark, which describes Ghostface Killah's alter ego and not Ironman. The words sword, flaming, and style are all connected to their debut album "Enter the Wu-Tang (36-Chamber)", which has a Shaolin theme.
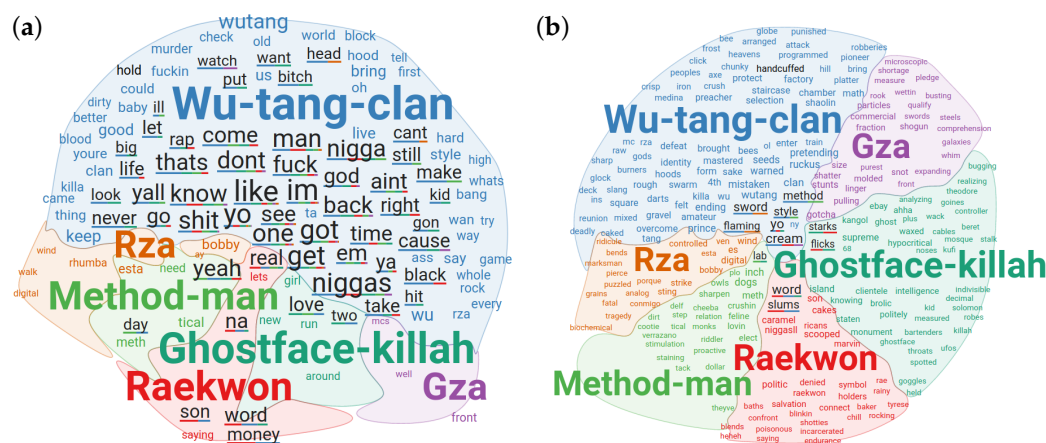

Figure 10. TagPie showing the most frequent used words by the Wu Tang Clan and various members (**a**). TagPie (**b**) shows words by z-score, e.g., words that are more frequent than the rest of the corpus.

4.9. Sentiment Analysis

Another facet of textual data is the sentiment it conveys. In order to communicate this facet, we computed a sentiment score for each line in the corpus between 1 (negative) and 5 (positive) For this, we used Huggingface [58] and the Multilingual Sentiment Analysis Bert Model by NLPTown [59]. With the sentiment score for each line, we computed an average sentiment score for each song and each artist in the corpus.

Design In the visualization system, a user can see a list of the German or English artists ordered by sentiment score either in ascending or descending order. Next to each artist, a colored rectangle based on the average score is displayed. When clicking on an artist of interest, an ordered list of the songs of the artist is displayed. For each song, a colored rectangle shows the average value for the song and a colored barcode shows the sentiment over the whole song for each line. The sentiment scores are mapped on a diverging color scale between red (negative) and blue (positive) with white for the neutral value.

Use Case On the left side in Figure 11 are the songs by American rapper 6ix9ine, in the middle are the songs by Macklemore and Ryan Lewis, and on the right side are the songs by the German rapper MCFitti ordered by average sentiment. The songs by MCFitti have a high sentiment on average, which reflects his more cheerful music, most of the songs of

6ix9ine have a low sentiment on average reflecting his aggressive music style, while the songs by Macklemore and Ryan Lewis range from positive to negative sentiment, showing a diversity from party songs such as "Can't Hold Us" and "And We Danced" and more serious themes such as drug addiction in "Otherside" and black lives matters and white privilege in "White Privilege II".
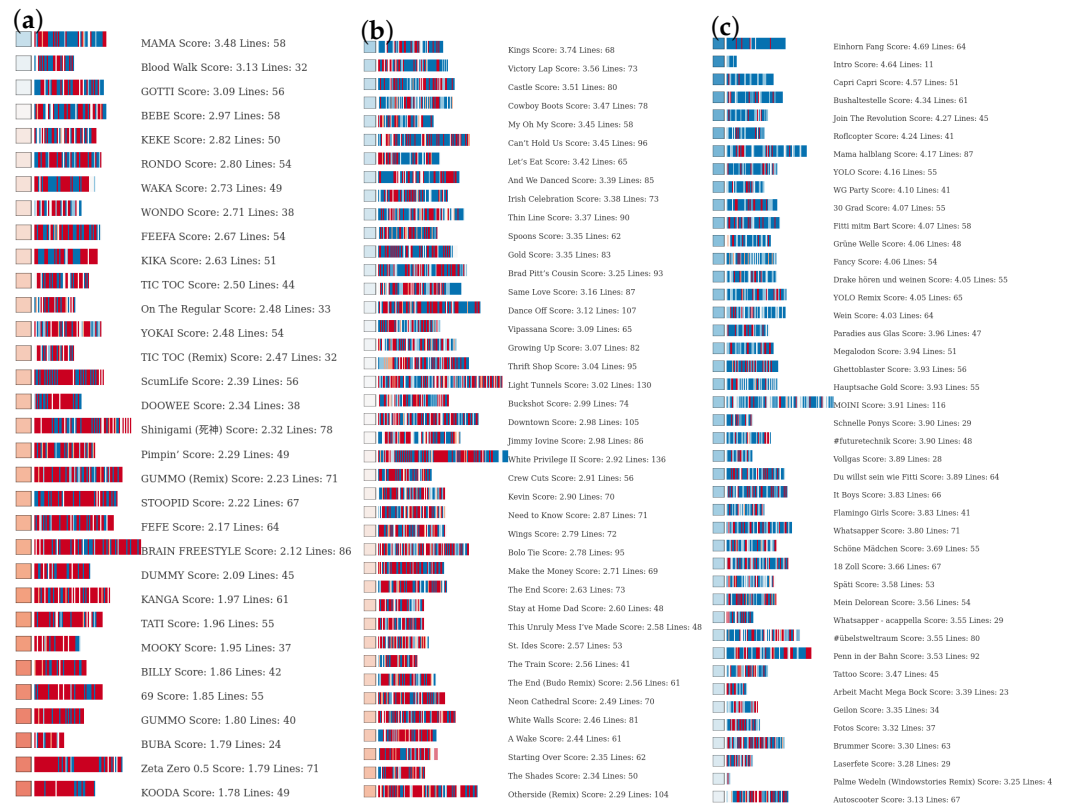

Figure 11. Sentiment Barcodes for the songs of 6ix9ine (**a**), Macklemore and Ryan Lewis (**b**), and MCFitti (**c**). A red bar indicates a negative sentiment, and a blue indicates a positive sentiment for a line.

5. User Feedback

We carried out an informal evaluation with six fans of rap music that have general and scene-specific knowledge about the German and US rap scene. They used the system for approximately half an hour up to one hour to explore the graph and the relations between the artists and the lyrics. One user suggested adding filtering by year for the song-level side-by-side view to focus on specific parts of the artist's career, e.g., when two artists were part of the same group, or if only early work or new work are similar. For example, he noticed a higher similarity in the lyrics of Tony D and Sido when both were part of the rap group "Die Sekte". A user noted that the list of similar songs in the profile view is helpful in detecting songs about the same or similar topics, e.g., love, cars, or drugs. Multiple users noted that the TagPies created by the z-score are helpful to confirm hypotheses about the vocabulary of two or more artists. For example, one user thought that the vocabulary of the Flatbush Zombies and The Underachivers is similar, which he then confirmed with the visualization. Users also noted that the relations in the graph make sense as long as the similarity value did not decrease too much.

6. Discussion

As in other digital humanities projects [60], the visualizations do not solely serve the purpose of delivering concrete answers to the research questions, and their purpose is more to generate new perspectives on the dataset and to trigger new hypotheses by allowing for

an exploratory analysis. In the following, we discuss inconsistencies in the data and the applied methods and potential directions for future work.

6.1. Imprecision and Incompleteness

A limitation of our approach is the data itself, and the data from Genius.com comprise different facets of inconsistencies [12], i.e., imprecision and incompleteness.

Although Genius.com always had a strong focus on rap music, there are probably always songs or artists that are not included and resulting therefore in an incomplete dataset. Furthermore, missing metadata information about artists or songs also leads to incompleteness. To increase the knowledge base, other information sources could be crawled and linked to the data from Genius.com. The data are also imprecise for multiple reasons. One reason is that the data are crowdsourced by users that use the website, thus resulting in typing errors in the lyrics or wrong artist information. Another inconsistency is given by the music genres, which are ambiguous terms without a clear starting point. Genre definitions or the association of a song or artists with a genre can change over time as new genres emerge. The visualization of temporal information about the genre tags can give an overview about the different types of rap genre and about new emerging genres but is not precise.

Imprecision is also given by applied machine learning methods such as sentiment analysis and word and sentence embedding. These methods are biased based on the data they were trained on. Another imprecision is given by the artist's similarity. Using the cosine similarity and including the ten nearest neighbors of each line influences the text alignments. Currently, alignments often occur because of the use of the same proper names such as the artist's names or cities, and the usage of the same adlibs. Therefore, including a threshold and other metrics could be helpful. Unfortunately, the Genius.com dataset has no ground truth so it is not possible to evaluate the quality of the alignments and the similarity metric. Nevertheless, the exploratory analysis with the visualizations allows for an informal evaluation through domain knowledge about artists relations. Additionally, alignments often occur between the hook or the refrain of two songs, so for future work, it would be better to treat them differently to focus more on less clear similarities. Another problem of the approach is that, frequently, a reference is created through metaphors, rhyme structures, or rearrangements of lines, which are hard to detect for automatic methods. Even including word vectors trained on Urban Dictionary does not tackle this issue.

6.2. Future Work

It is possible to extend this approach from monolingual lyrics to multilingual lyrics to detect cases where, for example, German artists reused passages from American artists. For this, as a proof of concept, we used the lyrics of around 20 international artists to find multilingual alignments between their lyrics and the lyrics of the German artists. We applied the pre-trained LASER [61] model for 93 different languages to create multilingual sentence embedding. The LASER encoder maps similar sentences of different languages to similar vectors and can be used without any additional fine-tuning. An alignment, in this case, can be seen as a translation. We found some initial results where the German artists communicated that they reused parts of English songs. Furthermore, the approach is expandable to all music genres and the whole Genius.com database with over 12 million lyrics. A possible future work would be therefore to use all of the data from Genius.com to detect multilingual references and to compare the similarity between songs based on their lyrics on a large scale through new Distant Reading methods. For example, with the visualization of alignments beyond the line level to inspect multiple texts at the same time or cross-line connections.

To extend the similarity analysis, the combination of lyrics and sound features is of interest. Similar to Yu et al. [34], sound features can be included next to the lyrics to create a multi-modal approach that includes similarities for example in mood, melody,

tempo, or rhythm. For this, the sampling information from crowdsourced websites such as WhoSampled.com [62] can be used to show more relations between songs and artists.

Another interesting approach would be to include other parts of cultural heritage such as literature to display the development of famous quotes such as "Each one teach one" over time, across music genres, and beyond music lyrics. Even more, a temporal visualization including historical events could give insights into how these events impacted the music.

The application of stylometry methods could be of interest to use frequencies of uncommon words such as the Burrows Delta [63] to find lyrics that are unusual for a given artist and that are more similar to the lyrics of another artist and thus can serve as an indicator for ghostwriting. Such ghostwriters are often not communicated to the audience: "the silent pens might sign confidentiality clauses, appear obliquely in the liner notes, or discuss their participation freely" [64].

7. Conclusions

We propose a prototype to compute the similarities of rap artists and to find intertextuality between monolingual song lyrics based on word embedding. The analysis is supported by visualizations to explore similarities between the lyrics of rap artists. The investigation of the lyrics is further supported by different views showing the metadata from Genius.com and visualizing similar songs or lyrics through stream graphs to find similar songs and to investigate monolingual alignments in their lyrics. Furthermore, we allow a multi-faceted exploratory analysis of the lyrics focusing on the sentiment of the songs, the vocabulary of the artists, and the development of rap genres, thus supporting multiple visual text analysis tasks on the Genius data. We explained the current limitations of the system, which we noticed through user studies. Furthermore, we laid out the possible directions to focus upon, such as finding multilingual alignments on a large corpus of song lyrics and cross-modality.

References

1. Kopano, B.N. Rap music as an extension of the Black rhetorical tradition: " Keepin'it real". *West. J. Black Stud.* **2002**, *26*, 204.
2. Spotify AB. Top Tracks 2019 Deutschland. 2008. Available online: https://open.spotify.com/playlist/37i9dQZF1DX4HROODZmf5u (accessed on 27 October 2021).
3. GMG Inc. 2014. Available online: https://genius.com/ (accessed on 27 October 2021).
4. YouTube LLC. YouTube. 2005. Available online: https://www.youtube.com (accessed on 27 October 2021).
5. Spotify AB. Spotify. 2008. Available online: https://www.spotify.com/ (accessed on 27 October 2021).
6. SoundCloud Limited. SoundCloud. 2007. Available online: https://soundcloud.com/ (accessed on 27 October 2021).
7. Yousef, T.; Janicke, S. A Survey of Text Alignment Visualization. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 1149–1159. [CrossRef]
8. Meinecke, C.; Jänicke, S. *Detecting Text Reuse and Similarities between Artists in Rap Music through Visualization*; OSF: Charlottesville, VA, USA, 2021.
9. Lim, D.; Benson, A.R. Expertise and Dynamics within Crowdsourced Musical Knowledge Curation: A Case Study of the Genius Platform. *arXiv* **2020**, arXiv:2006.08108.
10. Jänicke, S.; Franzini, G.; Cheema, M.F.; Scheuermann, G. *Visual Text Analysis in Digital Humanities*; Computer Graphics Forum; Wiley: Hoboken, NJ, USA, 2016.
11. Moretti, F. *Distant Reading*; Verso Books: Brooklyn, NY, USA, 2013.
12. Khulusi, R.; Kusnick, J.; Meinecke, C.; Gillmann, C.; Focht, J.; Jänicke, S. *A Survey on Visualizations for Musical Data*; Computer Graphics Forum; Wiley: Hoboken, NJ, USA, 2020.

13. Kim, J.H.; Tomasik, B.; Turnbull, D. Using Artist Similarity to Propagate Semantic Information. *ISMIR* **2009**, *9*, 375–380.
14. Schedl, M.; Hauger, D. Mining microblogs to infer music artist similarity and cultural listening patterns. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 877–886.
15. Schedl, M.; Knees, P.; Widmer, G. A web-based approach to assessing artist similarity using co-occurrences. In Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05), Riga, Latvia, 21–23 June 2005.
16. Jänicke, S.; Focht, J.; Scheuermann, G. Interactive visual profiling of musicians. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 200–209. [CrossRef] [PubMed]
17. Vavrille, F. LivePlasma. 2017. Available online: http://www.liveplasma.com/ (accessed on 27 October 2021).
18. Spotify AB. Spotify Artist Explorer. 2018. Available online: https://artist-explorer.glitch.me/ (accessed on 27 October 2021).
19. Gibney, M. Music-Map. 2011. Available online: https://www.music-map.de (accessed on 27 October 2021).
20. Cano, P.; Koppenberger, M. The emergence of complex network patterns in music artist networks. In Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR), Barcelona, Spain, 10–14 October 2004; pp. 466–469.
21. Gleich, M.D.; Zhukov, L.; Lang, K. The World of Music: SDP layout of high dimensional data. *Inf. Vis.* **2005**, *2005*, 100.
22. Daniels, M. The Largest Vocabulary in Hip Hop. 2014. Available online: https://pudding.cool/projects/vocabulary/ (accessed on 27 October 2021).
23. Schramm, K. Wer Hat den Größten? 2015. Available online: https://story.br.de/rapwortschatz/ (accessed on 27 October 2021).
24. The DataFace; Daniels, M. The Language of Hip Hop. 2017. Available online: https://pudding.cool/2017/09/hip-hop-words/ (accessed on 27 October 2021).
25. Lévesque, F.; Hurtut, T. MuzLink: Connected beeswarm timelines for visual analysis of musical adaptations and artist relationships. *Inf. Vis.* **2021**, *20*, 170–191. [CrossRef]
26. Lu, S.; Akred, J. History of Rock in 100 Songs. 2018. Available online: https://svds.com/rockandroll/#thebeatles (accessed on 27 October 2021).
27. Schedl, M.; Knees, P.; Widmer, G. Discovering and Visualizing Prototypical Artists by Web-Based Co-Occurrence Analysis. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK, 11–15 September 2005; pp. 21–28.
28. Logan, B.; Kositsky, A.; Moreno, P. Semantic analysis of song lyrics. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), Taipei, Taiwan, 27–30 June 2004; IEEE: New York, NY, USA, 2004; Volume 2, pp. 827–830.
29. Baumann, S.; Hummel, O. Using cultural metadata for artist recommendations. In Proceedings of the Third International Conference on WEB Delivering of Music, Leeds, UK, 15–17 September 2003; IEEE: New York, NY, USA, 2003; pp. 138–141.
30. Oramas, S.; Sordo, M.; Espinosa-Anke, L.; Serra, X. A semantic-based approach for artist similarity. In Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference, Malaga, Spain, 26–30 October 2015; pp. 100–106.
31. Knees, P.; Schedl, M. A survey of music similarity and recommendation from music context data. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2013**, *10*, 1–21. [CrossRef]
32. Deldjoo, Y.; Schedl, M.; Knees, P. Content-driven Music Recommendation: Evolution, State of the Art, and Challenges. *arXiv* **2021**, arXiv:2107.11803.
33. Ribeiro, R.P.; Almeida, M.A.; Silla Jr, C.N. The ethnic lyrics fetcher tool. *EURASIP J. Audio Speech Music Process.* **2014**, *2014*, 27. [CrossRef]
34. Yu, Y.; Tang, S.; Raposo, F.; Chen, L. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2019**, *15*, 1–16. [CrossRef]
35. Sasaki, S.; Yoshii, K.; Nakano, T.; Goto, M.; Morishima, S. LyricsRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics. In Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference, Taipei, Taiwan, 27–31 October 2014; pp. 585–590.
36. Ono, J.; Corrêa, D.; Ferreira, M.; Mello, R.; Nonato, L.G. Similarity graph: Visual exploration of song collections. In *SIBGRAPI*; IEEE, Institute of Electrical and Electronics Engineers United States: New York, NY, USA, 2015.
37. De Prisco, R.; Lettieri, N.; Malandrino, D.; Pirozzi, D.; Zaccagnino, G.; Zaccagnino, R. Visualization of music plagiarism: Analysis and evaluation. In Proceedings of the 2016 20th International Conference Information Visualisation (IV), Lisbon, Portugal, 19–22 July 2016; IEEE: New York, NY, USA, 2016; pp. 177–182.
38. Abdul-Rahman, A.; Roe, G.; Olsen, M.; Gladstone, C.; Whaling, R.; Cronk, N.; Morrissey, R.; Chen, M. *Constructive Visual Analytics for Text Similarity Detection*; Computer Graphics Forum; Wiley: Hoboken, NJ, USA, 2017; Volume 36, pp. 237–248.
39. Jänicke, S.; Geßner, A.; Büchler, M.; Scheuermann, G. Visualizations for Text Re-use. In Proceedings of the Information Visualization Theory and Applications (IVAPP), Lisbon, Portugal, 5–8 January 2014; IEEE: New York, NY, USA, 2014; pp. 59–70.
40. Asokarajan, B.; Etemadpour, R.; Abbas, J.; Huskey, S.J.; Weaver, C. *TexTile: A Pixel-Based Focus+ Context Tool For Analyzing Variants Across Multiple Text Scales*; EuroVis (Short Papers); The Eurographics Association: Norrkoping, Sweden, 2017; pp. 49–53.
41. Di Pietro, C.; Del Turco, R.R. Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. *Bleier Klug Neuber Schneider* **2018**, 133–163.
42. Riehmann, P.; Potthast, M.; Stein, B.; Froehlich, B. *Visual Assessment of Alleged Plagiarism Cases*; Computer Graphics Forum; Wiley: Hoboken, NJ, USA, 2015; Volume 34, pp. 61–70.

43. Jänicke, S.; Wrisley, D.J. Interactive visual alignment of medieval text versions. In Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), Phoenix, AZ, USA, 3–6 October 2017; IEEE: New York, NY, USA, 2017; pp. 127–138.
44. Meinecke, C.; Wrisley, D.; Janicke, S. Explaining Semi-Supervised Text Alignment through Visualization. *IEEE Trans. Vis. Comput. Graph.* **2021**. [CrossRef] [PubMed]
45. Jänicke, S.; Geßner, A.; Franzini, G.; Terras, M.; Mahony, S.; Scheuermann, G. TRAViz: A visualization for variant graphs. *Digit. Scholarsh. Humanit.* **2015**, *30*, i83–i99. [CrossRef]
46. Riehmann, P.; Gruendl, H.; Potthast, M.; Trenkmann, M.; Stein, B.; Froehlich, B. Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 1411–1423. [CrossRef] [PubMed]
47. Dekker, R.H.; Middell, G. Computer-supported collation with CollateX: managing textual variance in an environment with varying requirements. *Support. Digit. Humanit.* **2011**, *2*.
48. Brehmer, M.; Munzner, T. A multi-level typology of abstract visualization tasks. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2376–2385. [CrossRef]
49. Munzner, T. *Visualization Analysis and Design*; CRC Press: Boca Raton, FL, USA, 2014.
50. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
51. Wilson, S.; Magdy, W.; McGillivray, B.; Garimella, K.; Tyson, G. Urban dictionary embeddings for slang NLP applications. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11 May 2020; pp. 4764–4773.
52. Ethayarajh, K. Unsupervised random walk sentence embeddings: A strong but simple baseline. In Proceedings of The Third Workshop on Representation Learning for NLP, Melbourne, Australia, 20 July 2018; pp. 91–100.
53. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [CrossRef]
54. Box, G.E.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. (Methodol.)* **1964**, *26*, 211–243. [CrossRef]
55. Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings 1996 IEEE Symposium on Visual Languages, Washington, DC, USA, 3–6 September 1996; IEEE: New York, NY, USA, 1996; pp. 336–343.
56. Darville, J. Report: Three 6 Mafia Launch $6.45 Million Lawsuit against $Uicideboy$ over Samples. 2020. Available online: https://www.thefader.com/2020/09/08/report-three-6-mafia-launch-s645-million-lawsuit-against-suicideboys-over-samples (accessed on 27 October 2021).
57. Jänicke, S.; Blumenstein, J.; Rücker, M.; Zeckzer, D.; Scheuermann, G. TagPies: Comparative Visualization of Textual Data. In Proceedings of the Information Visualization Theory and Applications (IVAPP), Funchal, Portugal, 27–29 January 2018; IEEE: New York, NY, USA, 2018; pp. 40–51.
58. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; Association for Computational Linguistics: Vancouver, BC, Canada, 2020; pp. 38–45.
59. Town, N. Bert Base Multilingual Uncased Sentiment. 2020. Available online: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment (accessed on 27 October 2021).
60. Hinrichs, U.; Forlini, S.; Moynihan, B. Speculative practices: Utilizing infovis to explore untapped literary collections. *IEEE Trans. Vis. Comput. Graph.* **2015**, *22*, 429–438. [CrossRef] [PubMed]
61. Artetxe, M.; Schwenk, H. Margin-based parallel corpus mining with multilingual sentence embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3197–3203.
62. Limited, W. WhoSampled. 2008. Available online: https://www.whosampled.com/ (accessed on 27 October 2021).
63. Burrows, J. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Lit. Linguist. Comput.* **2002**, *17*, 267–287. [CrossRef]
64. Cameron, H. Diddy's Little Helpers. 2016. Available online: https://www.villagevoice.com/2006/11/14/diddys-little-helpers/ (accessed on 27 October 2021).