Heidelberg University
Institute of Computer Science

Project Report for the lecture
Data Science for Text Analytics

Hate, Discrimination & Racism in German Rap - A Text Analytics Approach

Johannes Sindlinger:	3729339, Computer and Data Science, M. Sc. johannes.sindlinger@stud.uni-heidelberg.de
Gal Lebel:	3679087, Computer Science, B. Sc. gal.lebel@stud.uni-heidelberg.de

Github-Repository:	`https://github.com/gsindlinger/IDSTA-Text-Miners`
Dataset:	Song Lyrics via Spotify API & Genius [13] (see description below)
Advisor:	Nicolas Reuter

Abstract

Rap music has become an essential part of pop culture worldwide and culture in general, also in Germany. German rap plays a major role in the German music scene, with the top three song artists of 2021 in Germany being rap artists. In addition to its popularity, German rap is known for its use of profanity and vulgarity, for its harsh lyrics and for the very sensitive topics it addresses. In this work, we wanted to investigate to what extent the accusations regarding the high level of violence, hatred and discrimination in German rap can be confirmed using methods of text analysis by machine. For this purpose, we collected data from 5992 songs of German rap artists in the years 1998 to 2022. We counted occurrences of specific terms of different categories of hate and discrimination, applied models of sentiment and toxicity analysis and also tried to assign songs to different classes via zero-shot classification. We were able to identify a tendency towards negative expressions in German rap, but at the same time it proved difficult to apply the methods of natural language processing. This was due to the fact that German rap songs usually do not follow the common syntactic scheme of German language.

1 Introduction

'I leave no whore daughter unfucked, everyone wants my dick - even lesbians get turned around!' - Excerpts from song lines by German rappers such as Bausa [26] provide material for discussion in German society and pose the question of how far artistic freedom can go in music and where insurmountable boundaries are crossed. Whether homophobia [26], misogyny [26] or antisemitism [23], in the public perception German rap seems to be one thing above all: Harsh and unfair. The popularity and sales figures of German rappers, on the other hand, justify their song texts and acting: at the end of October 2022, there were a total of ten titles in the top 20 singles charts in Germany that can be assigned to the genre of German rap [14]. And in 2021, rapper Capital Bra was the most successful German musician in terms of the number of different number 1 hits [6].

Contrary to the general negative impression, there are many attempts by artists who oppose against the negative image of rap in Germany with their lyrics and actions [27]. Some artists use their songs also used to specifically address sociopolitical issues - such as the 'Black Lives Matter' movement, police violence or the integration of refugees [20].

In this project, we would like to investigate the controversial debate around German Rap in an analytical manner. For this purpose, the song lyrics of various successful rappers of the genre of German rap will be analyzed on the basis of methods of textual data science. The following questions are the focus of our studies:

RQ1. **Do song lyrics of German rap in general possess a negative sentiment?**

RQ2. **Does hate, discrimination & racism exist in German rap song lyrics?**

RQ3. **How prevalent is hate, discrimination & racism in German rap song lyrics?**

Detailed ideas to answer these questions, including the data pipeline which we want to use, are described in section 3. Before that, the project will first be put into the context of existing literature in section 2.

2 Related Work

In 2018, the two known German Rappers 'Kollegah' and 'Farid Bang' have won a ECHO-Prize, despite their antisemitic text lines (see Feuerbach, [12]). This implies the significant and popularity of German rap in German society, despite its negative image and very aggressive nature. It is also worth to note, that Gangsta-Rap is very popular especially among young people and that it has been studied and found to have a negative influence on them [4, 3, 23]

Various journalistic and social science works in the past have dealt with the role of German rap in society.

The beginning of the 2000s marks a significant increase in the amount of German rap texts containing vulgarity, misogyny, sexism, anti-Semitism and violence. From about 4-5% of German rap songs containing sexist terms, the 2000s marked a jump towards ca. 25% of the songs containing such terms. Between 2005 and 2013 the trend has declined only to later on in 2018 go up again. An explanation for this might be, that sexism in rap songs has become more subtle by using less sexist terms but at the same time they still promote the sexist image and is also harder to detect by listeners as much as by means of text analysis. [22]

This general rise of this very violent/hatred-focused rap is directly connected to the rise of the 'gangster rap', which has become the most successful sub-genre of rap in general (like in the USA) and in German rap in particular. Gangster-Rap concentrates mostly on on the so called prison-culture. In the lyrics of such songs, one encounters very often terms related to violence, drugs, segregation from other (social) groups. It also conveys the hardships of being a minority in Germany and puts a spotlight on the socially weaker. [19]

In addition to sociotechnical analyses, there are two data-driven approaches to analyze the song lyrics of various German rappers. In 2016, Bayerischer Rundfunk's cultural magazine Puls [24] examined the political correctness of various song lyrics by German rappers, using a very similar methodology to the one we will use in this paper. Puls selected the five most commercially successful albums by German rappers in each year for the period 2006 to 2016 and downloaded the song lyrics via Genius. These song lyrics were examined for specific discriminatory word groups - with a particular focus on homophobic, racist, misogynistic, and ableist terms.

Puls observed that the use of discriminatory language increased over the first part of the sample period and decreased towards the end. Misogynistic and homophobic remarks played a particularly significant role. Discrimination against the disabled was also a permanent feature of the song lyrics studied, while racism was rather less prevalent. The author of the study also emphasizes the lower significance of the study due to the limitation to five albums per year.

In contrast to the analyses of Puls, we would like to get a broader view of the sentiment of German rap. Concretely, this means that we want to include data from more artists and songs in our analysis. In addition, we do not only want to consider frequencies of certain words, but more in-depth methods of text analysis, which are based on machine learning. Generally, the goal of this project is to gain as much information as possible about song lyrics and to determine their 'fairness' in social context. The insights of this project could also be extended to other music genres. In addition, the focus of this project is on German language, song lyrics in English could also be analyzed with the same approach.

3 Methods

As already outlined in section 1, we wanted to study the extent to which hate, discrimination and racism exist in German rap. In order to do so, we've used different methods of text analysis, which will be explained in more detail below. The described approach of our pipeline is also supplemented by a visual representation in Figure 1.

Our pipeline can be split into the following stages. Details regarding the individual points are detailed below:

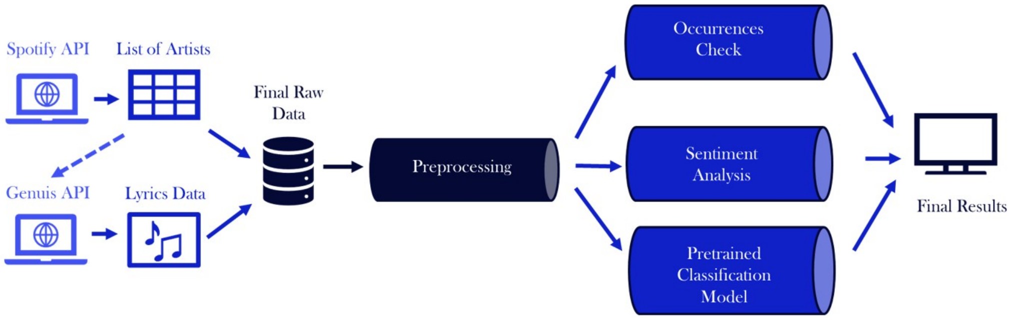

Figure 1: Text Analytics Pipeline

1. Data acquisition
2. Preprocessing
3. Text analysis methods: Occurrences check, Sentiment analysis, Zero-shot classification

Data acquisition

Since the subject of interest for our project was the analysis of German rap texts, we first needed to gather a dataset of German rap lyrics. We've decided to start with obtaining a list of German rap artists from the most popular German rap playlists on Spotify between the years 1998 and 2022. This was done with a Python library called spotipy [2], which works with the Spotify API.

For our analysis to have a meaningful interpretation of German rap's influence on German society, we decided to scrape the 15 most popular songs by each artist. In addition to the lyrics, we also saved relevant information that could be useful in later parts of the analysis. This information included the album to which a song belongs, the song's date of release, featured artists, etc.

To acquire this information, we've decided to use Genius [13], a website which provides a vast amount of lyrics for many songs. Genius offers their own API which allows for scraping of lyrics. In addition, the API has a Python library to work with. However, this API has different limitations regarding the amount of songs allowed to be scraped per artist, as well as the available information for the songs.

In order to overcome those limitations, we partly used the Genius API and the BeautifulSoup library [7]. This also allowed us to speed up the scraping process by using multi-threading, which wasn't possible with the API due to limitations on the amount of requests allowed at a given moment. The initial resulting dataset contained about 10,500 German songs from about 800 artists.

Preprocessing

After scraping, the resulting dataset needed to be preprocessed. Many of the songs we've scraped contained annotations for the different parts of the songs, such as the different verses and the chorus. These annotations sometimes also included the names of the artists in the case of having multiple of them in one song. Those parts were usually given by brackets, e.g. [Chorus] or [Verse 1], which made it easier to remove. Furthermore, many songs contained commas and points in places where they don't belong, or sentences that were cut by an additional line. We've used Python's Regular expression library (regex) to take care of the bracketed parts and the additional unwanted punctuation.

Upon inspection of the different scraped songs, we noticed that our dataset contained some songs for which the German language wasn't the major language. In fact, some of the songs didn't contain German words at all. For our analysis, it was crucial to filter out those songs. To do so, we could benefit from the Polyglot library [5], which helped us to remove undesirable songs. This step reduced our dataset to about 5992 songs.

For further preparation of the dataset for the analysis part, we had to get rid of stopwords and lemmatize the text. For the removal of stopwords, we were faced with several lists of stopwords for German language ranging from about 300 words to 1500+ words [15]. The different stopwords lists offered different advantages and disadvantages. For example, removing critical information versus keeping information that may not contribute or even undermine the success of the analysis. We've eventually decided to work with the stopwords list built by Gene Diaz [11], which has about 620 stopwords. For the lemmatization part, we've used the 'de_core_news_sm' model from nlp framework spaCy [18].

A substantial issue, which could have posed lots of implications and possibly prevented us from being able to correctly analyze the texts, was the lack of punctuation in the vast majority of the songs. We wanted to be able to break down the songs into meaningful sentences that could be processed and analyzed. To solve this issue, we've used a punctuation restoration library called 'Deep Multilingual Punctuation Prediction' [17]. This is a NLP model which was made for restoring punctuation of transcribed spoken language and was trained on the Europarl dataset [1].

Text analysis methods

Occurrences check

One big point of interest we wanted to investigate was the changes in trends and narratives over time in the German rap genre, as already discussed by Puls magazine and Spiegel magazine (see section 2).

We've decided to carry out an occurrence check on specific words in the songs and to roughly label songs using those occurrences to draw conclusions. Firstly, dictionaries with positive and negative sentiments were manually created for each of the following labels : Homophobia, misogyny, anti-disability, anti-semitism, racism, violence, love and grief. The last two categories were added to provide adequate comparability of the results with respect to the other categories.

Due to the fact that those dictionaries were manually created by us, we couldn't catch all possible synonyms and related words for each label. To overcome this, initial words were given for each category and synonyms were searched via Word2Vec model and its built-in k-nearest-neaghbour method. A pre-trained model by Andreas MÃ¼ller [21] served as a basis. This model is built on texts from the German pages on Wikipedia and German news platforms. This causes certain complications, as the base corpus of MÃ¼ller only partially corresponds to the language used in rap genre and particularly excludes vulgar words.

Accordingly, we had to manually adjust and correct the resulting dictionaries, which were selected on the basis of the 20 closest neighbours of the given input words.

In order to determine the specific occurrences of the given categories in the lyrics, all lyrics were checked against all dictionaries of the categories. We allowed substrings of the matching pairs in a certain way: If one of the words from the lyrics started or ended with a term from a category dictionary, this was counted as an occurrence. Likewise, an occurrence was counted if at least 75% of a word from the lyrics was identified as a substring in a corresponding word from a category dictionary.

Sentiment analysis

In addition to that counting analysis, we wanted to look for different pre-trained classification models to include the power of advanced natural language processing tools to enrich our analysis. To do this, we looked for models that could classify the sentiment of the songs in the context of our research questions. We found two different, BERT [10] based models which we could apply to our data: 'German Sentiment Bert' [16] by Oliver Guhr and 'German Toxicity Classifier Plus V2' [25] by Elisei Stakovskii.

The 'German Sentiment Bert' classification model was trained on 1.834 million German samples and outputs the probalities for the three classes 'negative', 'neutral', 'positive'. For every song, we predicted the probability of each line and mapped it from -1 (negative) to 1 (positive) where 0 is neutral. More specifically, for each line, the probability of the highest class was extracted and multiplied by the previously mentioned values -1 (negative), 1 (positive) and 0 (neutral). We subsequently summed up these values across the entire song and divided it by the number of lines to get the average sentiment of the total song.

The 'German Toxicity Classifier Plus V2' model was trained for toxicity labeling on the classes 'toxicity' and 'neutral'. Author Elisei Stakovskii built this model by fine-tuning the 'German BERT model' of MDZ Digital Library team at the Bavarian State Library. For the toxicity classifier, we processed the songs in the same manner as we did for the German Sentiment Bert model: we predicted the lines of each song, then summed the results up and averaged them. We again wanted to have a continuous mapping, which we defined as -1 to be not toxic and 1 to be toxic. For each line, we therefore took the most prominent class and multiplied it by -1 if it was 'neutral' and by 1 if it was 'toxic'.

Zero-shot classification

For this purpose, we decided to use a zero-shot classifier [9]. A zero-shot classifier is an NLP model that receives as input a sentence, a question about that sentence and a list of labels to which the model should assign the sentence depending on the input question. The model used is based on the XLM-RoBERTa model [8] and was then refined for the classification task. The model allowed us to create our own labels and predict the probabilities for a given sentence based on our own labels. For our task, it was sufficient to define the question "What is this example about?". At first we tried to use only the labels we were already interested in, namely 'homophobic', 'racist', 'misogynist', 'anti-semitic', 'violent', 'beautiful' and 'sad'.

However, we noticed that the model might misclassify a sentence if none of the labels fit exactly. This is due to the fact that it has no choice other than to assign the sentence to one of these labels. Therefore, in addition to our original labels, we decided to add some labels that can be used as offset labels, so the model can choose to use labels other than the labels we are interested in. Thus, the model does not always have to choose a label that is not necessarily the best fit. For example, in order to improve the quality of the classifier, we used the label "neutral" in this specific way. Finally, a sentence that is neither sad nor beautiful is not necessarily assigned to the other negative labels such as anti-semitic or misogynistic.

For the concrete classification of the songs, we followed the same approach as for the sentiment analysis. The preprocessed lines of each song were analysed via zero-shot classifier and the probabilities of the categories 'homophobic', 'racist', 'misogynist', 'anti-semitic', 'violent', 'beautiful', 'sad', 'lovely' and 'neutral' were determined. We took the average of these probabilities across all lines of each song. Thus we obtained a distribution for probabilities of the described categories of a whole song.

4 Results

As described in section 3, the data of 5992 songs by German rap artists formed the basis of our research. For each song, information on the artists contributing, the album and the date of release was stored. This information was obtained using the Genius API. In addition, the final data set contains the lyrics to the corresponding song, which were treated using the preprocessing procedure described above and also saved in modified form.

The 5992 songs that remained after discarding non-German songs include 2094 participating artists and are spread among 1711 different albums. The number of artists involved includes not only primary artists, but also producers and featured artists. For example, the two producers Tim Wilke and David Kraft are the most frequent artists with 68 occurrences in the 5992 songs. The third most frequent artist in the available lyrics is Sido with 65 occurrences.

888 songs were declared as singles by Genius and were therefore not assigned to an album. This class also forms by far the largest share of the existing albums. The other existing albums are largely based on the limitation of song scraping to 15 songs per artist (see section 3). The albums 'Instinkt' and 'Berlins Most Wanted' are both listed as 15 albums of one song. Only the album 'Liebeskummerparty' has 16 occurrences due to a song with a different artist.

Figure 2 shows the temporal distribution of the songs. As it can be seen from the figure, the data set contains considerably fewer songs in the years 1998 to 2010 than in the period from 2010 to 2022. This could be due to the procedure for generating the data, which is essentially based on the predefined playlists from Spotify and the availability of various songs on the lyrics platform Genius. In addition, the number of artists in the Deutschrap genre has grown steadily over the years and was very low when the genre first emerged. The described disparity of the data affects the interpretability of the analyses. Details on this are explained in detail in the following sections.

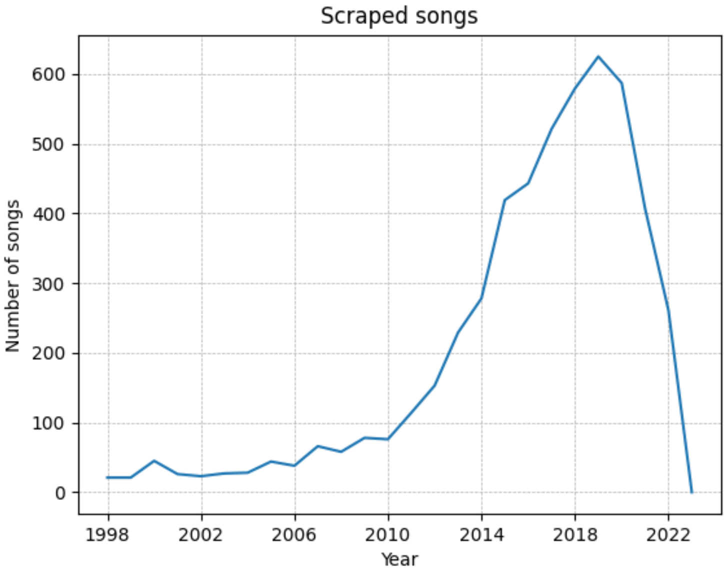

Figure 2: Temporal distribution of scraped songs

In the following, the results of the three different approaches to analyze the data will be presented in detail.

Occurrences check

As outlined in section 3, an essential part of the task of the occurrences check was to elaborate a suitable dictionary for different categories, on the basis of which occurrences within the lyrics could be retrieved. After executing the process described in section 3 via Word2Vec and manual adjustment, the following quantities of words remained per category:

The different number of words per category is due to the fact that certain categories have fewer diverse terms than others. For example, the category misogyny contains a wide variety of vulgar terms for prostitutes, while the racism category almost exclusively contains the word 'nigger'.

The following Table 2 shows the absolute count of occurrences of the different categories in the entire data set. The categories love, misogyny and violence combine more than 85% of all occurrences with 31451 of 36833 counted occurrences. 3628 of the examined songs contain at least one occurrence of the category love, which corresponds to about 61% of the entire

Category	**Number of words**
Misogyny	19
Violence	17
Anti-Semitism	14
Homophobia	14
Anti-disability	13
Grief	12
Love	10
Racism	6

Table 1: Number of terms within each category of occurrences check

data set. The category violence was detected at least once in 3099 songs, approximately 52% of all songs. We detected misogynistic terms in 2431 songs, about 41% of the population. In 757 songs, no occurrences of the predefined categories could be found.

Category	**Number of occurrences**
Love	12098
Violence	10251
Misogyny	9102
Racism	2051
Grief	1294
Homophobia	1253
Anti-disability	546
Anti-semitism	238

Table 2: Absolute count of occurrences of investigated categories

For further analysis, we examined the behaviour of the occurrences based on the release date of the songs. Due to the previously mentioned bias in the number of songs per year, we normalised the number of occurrences per year. Figure 3 shows the development of the occurrences over the examined period 1998 to 2022. The different lines show the normalised number of occurrences per song for each category. The different categories are marked with different colours. It can be observed that especially around the year 2003, a relatively large number of songs contained violent and misogynistic terms, but at the same time also many on the subject of love. Until 2010, the occurrences of the three mentioned categories decrease to an average level of 3 occurrences per song. They maintain this level in the following years up to 2022. The other categories are less prevalent throughout the entire study period: Only the categories homophobia in 2006 and racism in 2003 exceed the mark of 1 average occurrence per song.

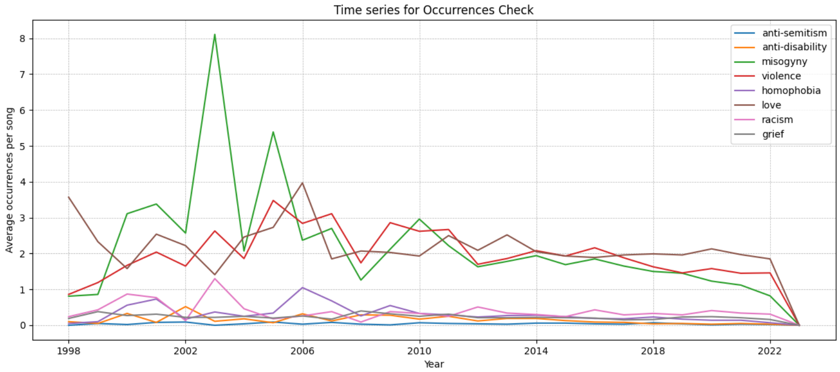

Figure 3: Time series analysis of counted occurrences

Sentiment analysis

For the analysis of the two described methods for sentiment analysis (German Sentiment Bert and Toxicity), we studied the distributions of the values across the data set. Figure 4 shows the distribution of the values of the German-Sentiment-Model according to the described procedure in section 3. Similarly, Figure 5 shows the distributions of the values of the toxicity classifier. The left graph of the figure represents a histogram of the distribution over the complete investigation period. The right graph in contrast shows the temporal course of the values. The middle blue line within this graph corresponds to the median of the examined songs within each year, the surrounding shaded area contains all data that falls into the range of the 25%- to 75%-quantile. In other words, 50% of the songs examined have a sentiment or toxicity value within the shaded area.

In general, one can observe for the sentiment analysis via German Sentiment Bert that a negative sentiment prevails in the songs. The median of all data is -0.40, the 25% quantile is -0.52 and the 75% quantile is -0.27. The clustering of data in this range can also be gathered from the histogram in Figure 4. Only 263 songs were assigned a sentiment greater than 0, which corresponds to a share of 4.4% of all the songs examined.

The time series analysis of the sentiment data on the right-hand side of Figure 4 reveals no meaningful change in the sentiment of the songs studied over the research period. The median decreases slightly from -0.36 in 1998 to -0.42 in 2022. The lowest median can be found in 2009 with a value of -0.47, and the highest median in 2003 with a value of -0.31.

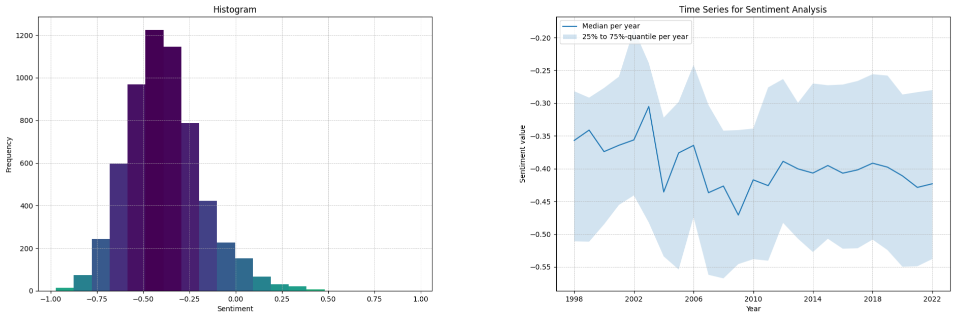

Figure 4: Distribution of results regarding German-Bert-Sentiment analysis

The toxicity analysis also shows a rather one-sided distribution of the examined data, as the histogram on the left side of Figure 5 shows. 1100 songs are classified with a positive value by the toxicity classifier, i.e. they are classified as rather toxic. This corresponds to a share of 18% of the population. The median of the toxicity values of all songs is -0.34, the 25% quantile is -0.60 and the 75% quantile is -0.08. The development over time, shown on the right side in Figure 5, shows a slight downward trend over the years. In other words, songs are becoming less toxic based on the toxicity analysis, but with the starting level in the 2000s already being classified as non-toxic.

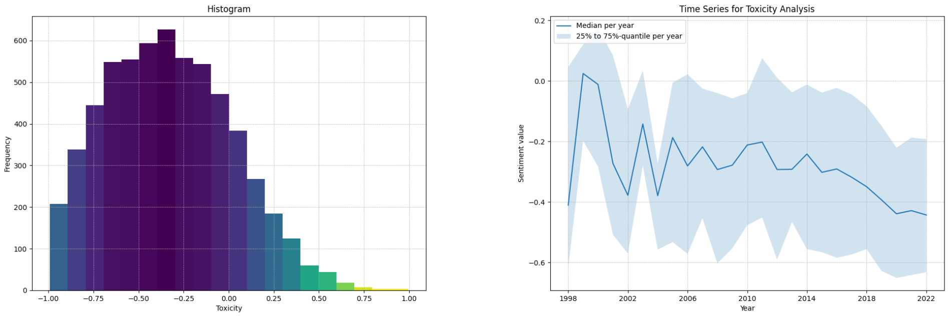

Figure 5: Distribution of results regarding Toxicity analysis

For a further analysis of the two methods of sentiment analysis, we compared the results of the two classifiers. Pearson's correlation coefficient provides insights into a possible dependency of the data. The result of the two variables German Sentiment Bert and Toxicity is -0.13.

Zero-shot classifier

The classification via zero-shot method (see section 3) provides probabilities for the given classes per song. Figure Figure 6 shows the distribution of classes across the data set. The graph on the left only considers the most likely class per song and ignores all probabilities of the other classes. It thus corresponds to a one-hot-encoded classification to a single category per song. As one can see from the graph, the class 'positive' is the most prevalent with 2149 songs assigned to it, followed by the category 'friendly' with 1732 occurrences and 'violent' with 935 occurrences. The categories 'affectionate', 'neutral' and 'mysoginistic' were selected as the most likely category only a few times. 'Racist' and 'homophobic' were not assigned to any song with highest probability.

Looking at the left side of the Figure 6, one gets insight into the summed probabilities of the categories. This reveals a slightly differentiated picture: The categories 'positive' and 'friendly' remain at the top, but with a smaller difference compared to the one-hot-encoded approach on the left side of Figure 6. The categories 'affectionate', 'violent', 'neutral' and 'mysoginistic' gain more importance in this view. Racist' and 'homophobic' are still relatively underrepresented.

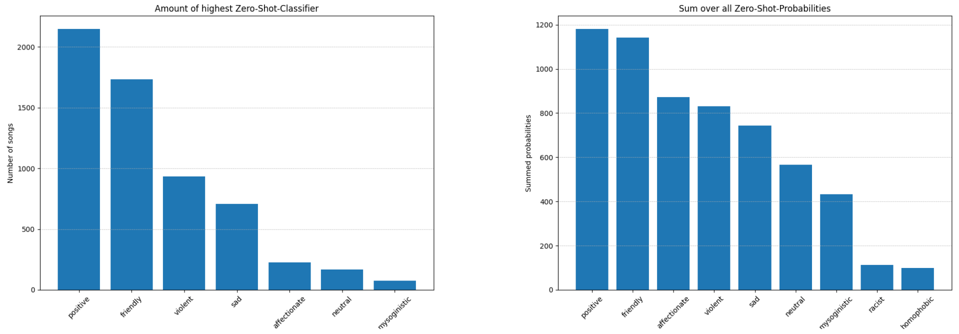

Figure 6: Distribution of categories regarding zero-shot classifier

The following diagram in Figure 7 provides an in-depth view of the development of the data over time with regard to the zero-shot classification: Four-year periods in the study period 1998 to 2022 are each shown with the three predominant categories in relation to the summed probabilities of the classifier. Due to the bias in the number of songs in the different periods (see Figure 2), the sum of the probabilities was divided by the number of songs per period. Over the period studied, the values of the two most probable categories ’friendly’ and ’positive’ are close to each other and continue to converge. Until 2014, the third most relevant category is ’violent’, before it is replaced by ’affectionate’. The average probability for the third highest category is between 12.5% and 15%.

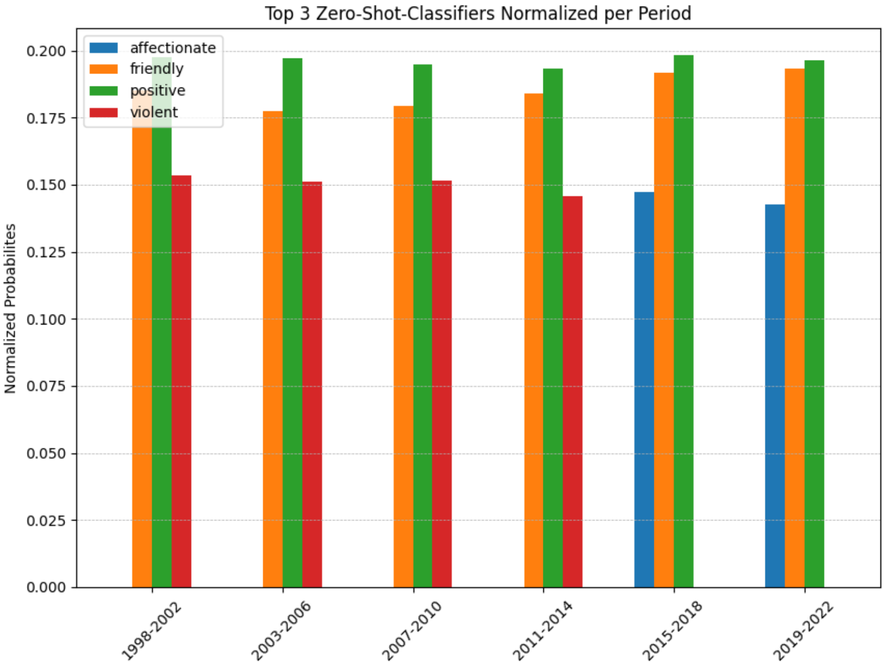

Figure 7: Time series analysis for the top 3 prevalent categories of zero-shot-classifier

Evaluation

In order to be able to classify the results and assess their quality, we have developed methods for evaluating the results. No evaluation was carried out for the occurrence test, as this method is based on pure counting and therefore only allows for misinterpretation to a limited extent. For the other methods, we faced the challenge that the data used was not labelled as it should be for a complete evaluation. Our dataset was self-acquired, so we had no data against which to compare the results of our methods.

One of the options we considered at this stage was to use ChatGPT OpenAI's Playground platform to automate the very time-consuming process of labelling the songs. Unfortunately, we encountered a few problems in the process. One of them was the inconsistency of the labelling, i.e. the same lines were interpreted and classified differently each time they are given as input to the AI. However, this might be something a human could do as well, one could interpret things from different perspectives at different times. One even more majore issue was the recognition of swear words in OpenAI, of which there are many in German rap, which completely ruled out this option for us.

We therefore decided to manually label about 70 songs, of which we had an equal number from each classes of the three different methods used, so we were able to equally evaluate the efficiency and accuracy of the methods in relation to the different labels we had. During the labelling process, we realised how difficult it is even for people to assign a label, be it a positive, negative or even more specific label (like racist, misogynistic, etc.). Many biases could influence the decision, and the songs leave a lot of room for different perspectives and interpretations of their content.

So the task of labeling the songs itself proved to be very challenging, which is why we decided not to give the chosen label a percentage probability, but to simply classify them in a binary way. So we decided to ignore the exact probabilities that the different methods assigned to the different possible classes and consider them as a binary classification.

For the evaluation, we've decided to use a Confusion Matrix and consider the F1 Score. We've compared our labels with the predictions of the different methods in the following manner:

- Sentiment Analysis (German Sentiment Bert): We mapped any negative sentiment, i.e. probability less than 0, to -1 and any positive sentiment to 1.

- Toxicity Analysis: We did the same approach as for German Sentiment Bert: Mapping was done to -1 for not toxic songs and 1 for toxic songs.

- Zero-shot Classification: We omitted the probabilities and took the class with the highest probability as the label of the song.

The confusion matrices for the different methods can be found below:

	Precision	Recall	F1-Score	support
negative	0.98	0.63	0.77	65
positive	0.08	0.67	0.14	3
accuracy			0.63	68
macro avg	0.53	0.65	0.45	68
weighted avg	0.94	0.63	0.74	68

Table 3: Confusion matrix for Sentiment Analysis (German Sentiment Bert)

	Precision	Recall	F1-Score	support
neutral	1.00	0.53	0.69	66
toxic	0.06	1.00	0.11	2
accuracy			0.54	68
macro avg	0.53	0.77	0.40	68
weighted avg	0.97	0.54	0.68	68

Table 4: Confusion matrix for Toxicity Analysis

	Precision	Recall	F1-Score	support
neutral	0.22	0.22	0.22	9
affectionate	0.78	0.70	0.74	10
violent	0.40	0.67	0.50	0.9
racist	0.00	0.00	0.00	0
homophobic	0.00	0.00	0.00	0
misogynist	0.60	0.33	0.43	9
friendly	0.00	0.00	0.00	10
positive	0.50	0.36	0.42	11
sad	0.48	1.00	0.65	10
accuracy			0.47	68
macro avg	0.37	0.41	0.37	68
weighted avg	0.43	0.47	0.42	68

Table 5: Confusion matrix for Zero-shot Classifier

5 Analysis

In the following, the results listed in section 4 will be put into the context of the research questions and will be interpreted. Furthermore, limitations of the project and possible extensions will be mentioned.

The analysis of the Occurrences Check shows that particularly violence and misogyny are an existing problem in the genre of German rap. As Figure 3 shows, occurrences of this category are present throughout the entire period of the study. This proposition is further supported by the fact that 51% of all the songs studied contain violent terms in some form. However, at the same time it must be taken into account that the category love has a similar behaviour and therefore represents an antithesis to the problem described. 61% of the songs studied have occurrences on the theme of love. Overall, these observations can be interpreted in the sense that emotionality plays a large role in German rap.

Furthermore, the analysis of the time series of occurrences check shows a slight decrease in the most prevalent categories misogyny, violence and mysogyny from 2005 onwards (see Figure 3). We suppose that these changes in German rap are related to transformations within German society, i.e. changes in terms of gender equality due to movements like MeToo-discussion, open-mindedness and acceptance towards gay community or changes in attitudes towards minorities. These transformations are not only due to evolving politics, but can also be seen internationally and especially in the US, which has a huge international influence on rap in general. As society adopts these attitudes and societal values, German rap artists might also become aware of the importance of their lyrics and the message they convey, which requires artists to adapt to these social changes. Another possible reason would be the constantly evolving vocabulary and slang, which could mean that our dictionaries and the words they contain fit better in the early 2000s than in later times.

The application of the sentiment analysis via the German Sentiment Bert and Toxicity-Model allows only limited interpretative space than the occurrences check due to relatively weak evaluation results. Generally, the two models show a strong one-sided allocation, which is shown both in the observation of the distributions via histogram (see Figure 4, Figure 5) and in the confusion matrices (see Table 3, Figure 5). Of the 70 songs evaluated, 26 were manually classified as positive, but the classifier recognised only 2, respectively 8%, as such. Similarly, the recall for the toxicity classification is 6% for the classification 'toxic', since only very few songs were actually classified as such by the classifier. Because of this bias, the values for the recall of the respective contrary class are very high. The precision is also of limited significance due to the fact that only two songs were classified as toxic and three as positive. In particular, the value 1.0 for the category neutral in the toxicity classifier is therefore meaningless. Overall, the rather sobering results of the evaluation can be summed up by the low macro average: This is 0.45 for the German Sentiment Bert and 0.45 for the Toxicity classifier. An improved evaluation that takes into account more data from the respective under-represented class could provide improved insights.

Regardless of the results of the evaluation, at least the bias of the data of the sentiment analysis of the songs towards rather negative sentiment could be explained to a certain extent. As mentioned before, emotions very often play a major role in songs, which are formulated in a particularly expressive way compared to natural language - whether on the topic of love, hate, etc.. This could cause the sentiment model to increasingly detect emotions and then evaluate them as negative, as this is frequently found in ordinary language use.

Comparing the results of German Sentiment Bert and the Toxicity model, we obtained a negative Pearson correlation of -0.13 (see section 4), which means that there is little to no noticeable correlation between the two models. This means that at least some songs were classified as negative but not toxic or positive but toxic. These results are in line with our expectations and raise interesting questions: 'What leads to the decision to rate a song as positive or negative?' 'Does the use of swear words mean that the message conveyed is necessarily toxic?'

It is possible to convey a positive message but use offensive words, while it is also possible to make a negative statement about something without being toxic. This is one of the difficulties with sentiment analysis and the way humans or a machine can interpret a text: Many factors can influence the decision, e.g. social prejudices, personal experience, gender, etc. In terms of natural language processing models, this means that the dataset on which the model has been trained has a big impact on the interpretation obtained by the model.

The results of the zero-shot classification should also be treated with caution due to imprecise evaluation results (see Table 5). However, it must be taken into account that the classification of eight different labels is significantly more difficult than for binary models. Accordingly, the weighted average of the F1-score of 0.52 is not ideal, but also not completely poor. The macro average of the F1-score, however, is rather weak at 0.37. Apart from the classes 'homophobic' and 'racist', which the model did not see as the most probable in any of the 5992 songs examined, the zero-shot classifier at least partially recognises the identical labels as we humans did. The absence of the classes 'homophobic' and 'racist' can be explained by a correlation of the classes to the category 'violent', which was always considered more relevant. Over all categories, the classification into 'affectionate' and 'sad' was the most accurate among all classes.

Based on Figure 6 and Figure 7, it can only be stated that violence at least does not seem to be completely irrelevant in the songs studied. The high proportion of songs classified as positive and friendly might not correspond entirely to reality, taking into account the various occurrences of terms with negative connotations in the context of the occurrences check.

6 Conclusion

During our work on the project, we encountered different difficulties that posed limitations to our ability to process the songs correctly. One of the main issues we dealt with was the way the texts are written. Due to the nature of the language used very often in German rap, such as slang, juvenile words, foreign words, and the texts being written and spelled phonetically and not the way they should be spelled, preprocessing those texts by means of NLP (such as lemmatization, tokenization, which all relay heavily on and follows a set of grammatical rules) was very difficult. For example, many words are spelled phonetically, e.g. in the way they're being spoken. An example of such a case is - ich hab' instead of ich habe. Another example is eine spelled as 'ne. In other cases, nouns weren't written with a capital letter, which made it hard to distinguish whether it is a verb or a noun. Due to those deviations from how the words should be written, lemmatization or the removal of stopwords will fail to work. In such a case, the word 'ne wouldn't be removed because it's not being detected as a stopword. We've tried to get around this issue by using regex, but it was impossible to address all the different cases. During our attempts we've noticed that words, that shouldn't be influenced by the different regex rules that we tried to use, were changed, thus rendering this option completely useless. This made it impossible, for example, to correctly learn the vocabulary of those songs using methods like Word2Vec. The same word, depending on how it is written, was treated as 2 different words (for example, in the case of hab' and habe, those were considered to be two different words). Another issue that could've potentially limited our results, even more, was the lack of punctuation. For correct processing of the texts, entire songs should be correctly split up into different sentences. The lack of punctuation, in this case, prevented us from being able to understand what word belongs to which part of the text. We were able to get solve this issue with the auto-punctuator model mentioned in the pipeline part. Without this solution, it would've been very difficult to process the songs even with pre-trained models.

References

[1] europarl dataset. https://huggingface.co/datasets/wmt/europarl.

[2] lightweight Python library for the Spotify Web API. https://spotipy.readthedocs.io/en/2.22.1.

[3] Antisemitism in gangsta-rap, 2020. https://www.uni-bielefeld.de/fakultaeten/erziehungswissenschaft/zpi/projekte/antisemitismus-gangsta-rap/.

[4] Wie gangsta-rap jugendliche hoerer beeinflusst. *Westfallen-Blatt*, 2021. https://www.faz.net/aktuell/gesellschaft/menschen/tv-kritik-zum-echo-2018-kollegah-und-farid-bang-gewinnen-echo-15539696.html.

[5] Rami Al-Rfou. Polyglot. https://polyglot.readthedocs.io/en/latest/.

[6] Bayerischer Rundfunk. Respekt: Deutschrap - erfolgreich gegen Diskriminierung?, Oct 2019. https://www.br.de/extra/respekt/deutschrap-diskriminierung-minderheit-100.html.

[7] Beautiful Soup. Beautiful Soup. https://beautiful-soup-4.readthedocs.io/.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. https://huggingface.co/xlm-roberta-large.

[9] Joe Davison. Zero-Shot Classifier based on xlm-roberta-large. https://huggingface.co/joeddav/xlm-roberta-large-xnli.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. http://arxiv.org/abs/1810.04805.

[11] Gene Diaz. The most comprehensive collection of stopwords for the german language. https://github.com/stopwords-iso/stopwords-de.

[12] Leonie Feuerbach. Provokation im rap muss grenzen haben. *FAZ*, 2018. https://www.faz.net/aktuell/gesellschaft/menschen/tv-kritik-zum-echo-2018-kollegah-und-farid-bang-gewinnen-echo-15539696.html.

[13] Genius. Genius - Song Lyrics & Knowledge. https://genius.com/.

[14] MTV Germany. Offizielle Single Top 100 - Musik Charts, Oct 2022. https://www.mtv.de/info/tyk12u/single-top100.

[15] Marco Götze. Plain and Extended list of German stopwords. https://github.com/solariz/german_stopwords.

[16] Oliver Guhr. German Sentiment Classification with Bert . https://huggingface.co/oliverguhr/german-sentiment-bert.

[17] Oliver Guhr. Deep Multilingual Punctuation Prediction. https://pypi.org/project/deepmultilingualpunctuation.

[18] Matthew Honnibal and Ines Montani. spaCy: Industrial-Strength Natural Language Processing. https://spacy.io/.

[19] Michael Huber. Gangsta-rap - wie soll man das verstehen? *BPJM-Aktuell*, 2018. https://www.bzkj.de/resource/blob/128956/a443538e6df7b950b4705bf83748f8ba/201803-gangstarap-data.pdf.

[20] ME-Redaktion. 5 Deutschrap-Songs Ã̈¼ber Alltagsrassismus, die wir kennen sollten, Feb 2021. https://www.musikexpress.de/5-deutschrap-songs-ueber-alltagsrassismus-die-wir-kennen-sollten-1824489/.

[21] Andreas MÃ¼ller. Analyse von Wort-Vektoren deutscher Textkorpora, June 2015.

[22] Bjoern Rohwer. Sexismus im Deutschrap: Wir haben 30.000 songtexte aus vier Jahrzehnten analysiert, Jul 2020. https://www.spiegel.de/kultur/musik/sexismus-im-deutsch-rap-text-analyse-aus-vier-jahrzehnten-rap-geschichte-a-8777bc4f-0c5d-461e-8d19-e99d69a3e3d0.

[23] Ben Salomo and Ludwig Greven. 'In der Rap-Szene existiert ein judenfeindliches Grundrauschen', Nov 2021. https://www.kulturrat.de/themen/texte-zur-kulturpolitik/in-der-rap-szene-existiert-ein-judenfeindliches-grundrauschen/.

[24] Matthias Scherer. Diskriminierende Texte: So politisch korrekt ist Deutschrap, Sep 2016. https://www.br.de/puls/musik/so-homophob-frauenfeindlich-rassistisch-und-behindertenfeindlich-ist-deutschrap-100.html.

[25] Elisei Stakovskii. German Toxicity Classifier Plus V2 . https://huggingface.co/EIStakovskii/german_toxicity_classifier_plus_v2.

[26] Friedrich Steffes-lay. Bausa sorgt auf 'Vossi bop' für einen der ekligsten Deutschrap-Momente des Jahres, Jul 2019. https://www.musikexpress.de/bausa-sorgt-auf-vossi-bop-fuer-einen-der-ekligsten-deutschrap-momente-des-jahres-1313477/.

[27] Tooka Tajali-Awal. Feminismus im Deutschrap - Paradox und Vielfältig, Dec 2021. https://www.deutschlandfunkkultur.de/hass-frau-paradoxer-feminismus-und-feministische-vielfalt-im-deutschrap-dlf-kultur-57d5044e-100.html.

Project contributions

Incl. former team members Mara-Eliana Popescu and Simon Körner

Date	**Who?**	**What?**
Oct 28	Johannes Sindlinger	Design pipeline graphics
Oct 28	Johannes Sindlinger	Write motivation, research topic (partly), project description (partly)
Oct 30	Mara-Eliana Popescu	Extend project description
Oct 31	Gal Lebel	Extend research topic, finish proposal
Nov 18	Johannes Sindlinger	Setup issues for project start
Nov 25	Johannes Sindlinger	Create artist list via Spotify API
Dec 10	Gal Lebel	Genius Lyrics Scraper
Dec 10	Gal Lebel	Milestone Editing
Dec 27/28	Johannes Sindlinger	Elasticsearch Connection
Jan 2	Gal Lebel	Preprocessing - cleaning up dataset
Jan 2	Gal Lebel	Filtering out non-German songs (polyglot)
Jan 12/13	Johannes Sindlinger	Category Dictionary via Word2Vec
Jan 13	Johannes Sindlinger	Counting Occurrences and Lemmatization
Jan 17	Gal Lebel	Auto-Punctuation for meaningful splitting into sentences
Jan 29	Gal Lebel	Zero-Shot classification
Feb 1 - 28	Johannes Sindlinger	Frontend Design
Feb 14/15	Johannes Sindlinger	Fastapi Service
Mar 01	Johannes Sindlinger & Gal Lebel	Manual Labeling & Evaluation

Declaration of Honour

We hereby declare on our honour that we have prepared this work independently; thoughts taken directly or indirectly from outside sources are marked as such. The work has not been submitted to any other examination authority and has not yet been published.

We are aware that an untrue statement will have legal consequences.

Heidelberg, March 10, 2023