Heidelberg University
Institute of Computer Science

Project Proposal 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
Mara-Eliana Popescu:	4166979, Computer Science, B. Sc. mara-eliana.popescu@stud.uni-heidelberg.de
Simon Körner:	3310142, Computer Science, B. Sc. simon.koerner@stud.uni-heidelberg.de
Team Member:	Name, Matriculation Number, Course of Study email address

1 Motivation

'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 [17] 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 [17], misogyny [17] or antisemitism [15], 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 [7]. And in 2021, rapper Capital Bra was the most successful German musician in terms of the number of different number 1 hits [4].

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 [18]. 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 [13].

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 Research Topic Summary

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

TODO: Summarize and search for research projects in this field! These are only some suggestions I could find. [9], [2], [20] journalistic: [14]

In addition to sociotechnical analyses, there is one data-driven approach to analyze the song lyrics of various German rappers. In 2016, Bayerischer Rundfunk's cultural magazine Puls [16] 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 Project Description

As already outlined in section 1, this project will study the extent to which hate, discrimination, and racism influence German rap. For this purpose, we would like to use different methods of text analysis, which will be explained in more detail below. The described approach will also be supplemented by a visual representation in Figure 1.

As a basis for our investigations, a data set on certain artists of the German Rap will be used, which is not finally defined yet. This will be done in the nearest future using information from the web in manual and

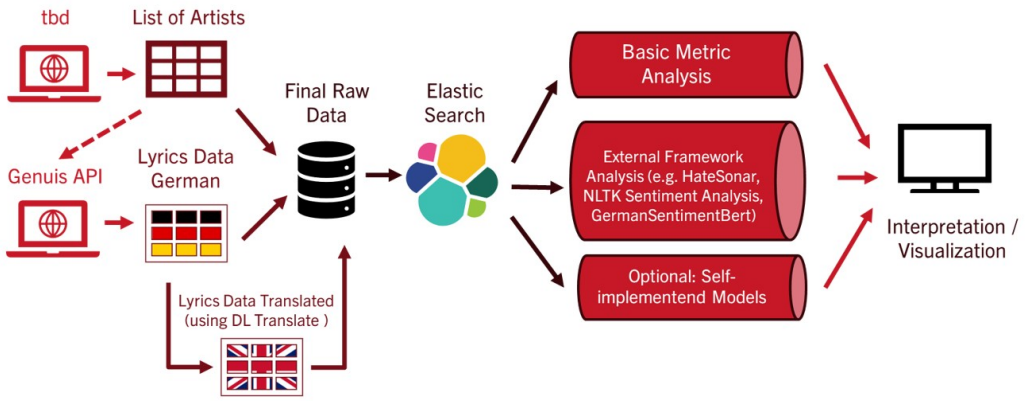

Figure 1: Text Analytics Pipeline

partially automated work. Corresponding information about artists can be extracted, for example, from [11, 19]. The names of those artists are used in a further step to download the lyrics of all songs of those artists via the song lyrics platform Genius [6]. Genius is an online database for any kind of artistic texts and offers an API interface that provides license-free lyrics of many song lyrics. Since Genius does not fully provide all the data of every international artist, it might be necessary to accept limitations for some artists.

The song lyrics, enriched with information about the artists themselves, finally form the basis of our textual analyses: Since we would like to use pretrained models for the analysis of the lyrics in a further step and many of these models only support English-language texts, it might be necessary to translate the German song lyrics first. For this, we intend to use the Python framework DL Translate [12], since it is the only package that can freely translate unlimited texts. The translated lyrics, will be stored together with the original lyrics in ElasticSearch. It must be taken into account that by using such a translation tool, possible linguistic-relevant contexts will be incorrectly transferred into the English language. However, since there is considerable interpretative space in the context of song lyrics, this limitation should not be of too much importance. In fact, it is important to keep in mind for the entire project that the song lyrics studied allow for different interpretations, which can only be determined by machine analysis to a limited extent.

As indicated in the paragraph before, we would like to store the translated song lyrics together with the original data in ElasticSearch. The possibilities that ElasticSearch offers in the area of tokenization, classification of words

including counting of word frequencies, etc. shall then be used as a basis for a first analysis of the song lyrics.

In addition to the described methods offered by ElasticSearch, we would like to use predefined machine learning models to enrich the data of the song lyrics. In this context, the frameworks used should be considered as a black box, and the corresponding methods remain untouched. We will consider the following frameworks:

The project Deep Learning Models for Multilingual Hate Speech Detection [3] includes multiple models that can be trained or fine-tuned for recognizing hate speech in various languages, especially German. Performing hate detection using different machine learning algorithms in parallel would enable a more thorough analysis of the texts like: highlighting songs that get a high hate rate from most models, identifying common patterns between such songs and classifying them into subclasses according to the target of the hate speech (foreigners, women, disabled people etc.).

The German Sentiment classification with BERT [8] is a sentiment classification model trained on around 1.8 million German-language text samples coming from various sources (social media, movie, app and hotel reviews). It could be a starting point for identifying potential hateful song lyrics. Given the three sentiment classes used by this model (negative, neutral and positive), we could filter out texts classified as positive and also separate negative from neutral lyrics for the upcoming pipeline stages.

NLTK.Vader [10] is an NLP algorithm trained for performing sentiment analysis. It is best suited for short texts like posts on social media, containing some slang and abbreviations. Hence, our dataset on German song lyrics would make a good fit for this model (providing we first use a translation model like mBart [1]).

HateSonar [5] is a BERT based model built with Python for hate speech detection. It only works with English data, therefore the German song lyrics would first need to be translated. As with models trained on German texts, it would be beneficial to the analysis to extract information regarding hate speech from multiple models.

If there is time left, we would like to develop independent machine learning models on our own using the methods discussed in the lecture. However, this requires classification of the existing lyrics data, which would likely need to be done manually. Furthermore, the amount of data available could prove challenging with this approach. If it is not feasible to identify a very large number of song lyrics using the listed approach above, it could be difficult to develop meaningful machine learning models.

Finally, the results of the different analysis methods will be interpreted and visualized. Thereby, the findings of the analyses shall be explicitly high-

lighted using the lyrics data by visual markers. Additionally, results of the different metrics will be displayed.

References

[1] Multilingual BART. https://huggingface.co/facebook/mbart-large-cc25?text=My+name+is+Sarah+and+I+live+in+London.

[2] Michael Ahlers. 'Kollegah the Boss': A case study of persona, types of capital, and virtuosity in German gangsta rap. *Popular Music*, 38(3):457–480, 2019.

[3] Sai Saketh Aluru, Punyajoy Saha, and Binny Mathew. Deep Learning Models for Multilingual Hate Speech Detection, Jun 2021. https://github.com/hate-alert/DE-LIMIT.

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

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017. https://github.com/Hironsan/HateSonar.

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

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

[8] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1627–1632, 2020. https://huggingface.co/oliverguhr/german-sentiment-bert?text=Das+ist+gar+nicht+mal+so+schlecht.

[9] Melanie Heinisch. *Schlampe, Hure oder Heilige. Das Frauenbild im Deutschrap*. GRIN Verlag, 2018.

[10] Clayton J. Hutto, Ewan Klein, Pierpaolo Pantone, George Berry, and Malavika Suresh. NLTK.Vader. https://www.nltk.org/_modules/nltk/sentiment/vader.html.

[11] Last.fm. Topkünstler von Deutschrap. https://www.last.fm/de/tag/deutschrap/artists.

[12] Xing Han Lu. DL-translate: A deep learning-based translation library built on Huggingface Transformers, Jul 2022. https://github.com/xhluca/dl-translate.

[13] ME-Redaktion. 5 Deutschrap-Songs Ä¼ber Alltagsrassismus, die wir kennen sollten, Feb 2021. https://www.musikexpress.de/5-deutschrap-songs-ueber-alltagsrassismus-die-wir-kennen-sollten-1824489/.

[14] 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.

[15] 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/.

[16] 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.

[17] 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/.

[18] 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.

[19] Tonspion. Die 50 Wichtigsten Deutschen Rapper, Jul 2021. https://www.tonspion.de/news/die-wichtigsten-deutschen-rapper.

[20] Martin Wiegel. *Deutscher Rap - Eine Kunstform als Manifestation von Gewalt?* Tectum Wissenschaftsverlag, 2011.