The Evolution of German Rap Lyrics

Luca Schumann
lucsc442
TDDE16

***Abstract*—Test**

I. Introduction

Introduce the task or research question that you have addressed in your project. What were you trying to do? Why did you choose this project?

II. Theory

Present relevant theoretical background, and in particular the models that you have used. Where appropriate, use mathematical formulas.

III. Data

Present your data. What information does it contain? Where did you get it from? What preprocessing did you do, if any?

A. Dataset

First of all, a selection of songs had to be made. To find a balance between relevant artists in the CD era and the streaming era, I decided to combine lists of artists from two sources. I selected all 46 german artists from the Wikipedia list of Rappers that sold more than 100,000 albums in Germany [1]. The popularity of musicians today, especially Rappers, is not only defined by the albums sales but also the online streams. Therefore I picked all 13 Rappers appearing in the Spotify playlist ”Top German Artists of 2020” [2] that were not in the selection yet. To increase the dataset, I chose 18 more objectively popular Rappers, adding up to 77 artists in the end.
The lyrics of the artists' songs were then downloaded using the Genius API [3] with help of the *lyricsgenius*package [4]. A lot of the songs are duplicates, since remixes, features and live versions are all included in the genius database. There are furthermore many non-songs in the database, such as commentaries. Luckily, most could easily be ignored by discarding songs that contain brackets, as in *Papa ist da (AsadJohn Remix)*. Few others had to be handled differently, but the final dataset is clean of such occurences. In a first run of the model, songs containing english lyrics formed an individual topic. Upon manual inspection, there were a few songs with parts in english, or even completely in english. Therefore, songs that contain all the most common english words in the english topic ”the”, ”you” and ”and” were suspected to contain english lyrics and also discarded. The final dataset features almost 7500 songs.

B. Preprocessing

Preprocessing the song lyrics was an iterative process with many learned lessons. The most basic first step is to remove meta information that is written in brackets from the lyrics, such as *[Hook]*. Next, the songs are tokenized using spacy's German package. Non-alphabetic tokens, stop words and tokens with less than three letters are removed. The rest of the preprocessing is done with Gensim. Bigrams and Trigrams are added and extremes are filtered. Finding the seemingly best values for these tasks was done with the trial and error method.
After excluding english songs from the dataset, the model would still find a topic with mostly english stop words. After removing english stop words during preprocessing, the problem was solved. Another problem was many artist's names appearing in the most salient words of some topics. The reason for this is that some Rappers are notorious for talking and singing about themselves in the third person. Rappers with a lot of songs in the dataset would therefore skew the results, although their own name has nothing to do with the message behind their song. To tackle the problem, each artists name was removed from his or her lyrics. As a final tweak of the preprocessing step, a couple of stop words that were not in spacy's German list were manually removed.
Tokenizing German text turned out to be a difficult task, with no really satisfying solution. In German texts, all nouns are capitalized, therefore also in spacy's word list. If a word at the beginning of a sentence is capitalized, it is not possible for spacy to tell wether it's a noun or not. Feeding spacy all words in lowercase would also not solve the problem, since it would not recognize the nouns anymore. Additionally, some words have a different meaning depending on wether they are a noun (capitalized) or not. E.g. ”Weg” means path but ”weg” means away in German. This appears to be an unsolved problem as of today, as can be read on spacy's Github forum [5]. Lemmatization does also not work as well as for english text, but overall the results are good enough to work with them.

IV. Method

Explain how you carried out your study. Aim to be detailed enough for others to reproduce your results.

V. Results

Present your results in an objective way. Use tables and charts, but do not forget to also include a summary in text form. Do not interpret your results

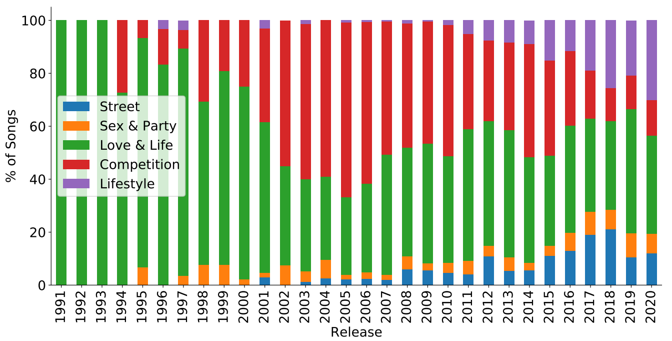

Fig. 1. Normalized timeline of the average topic distribution in the lyrics of songs released between 1991 and 2020.

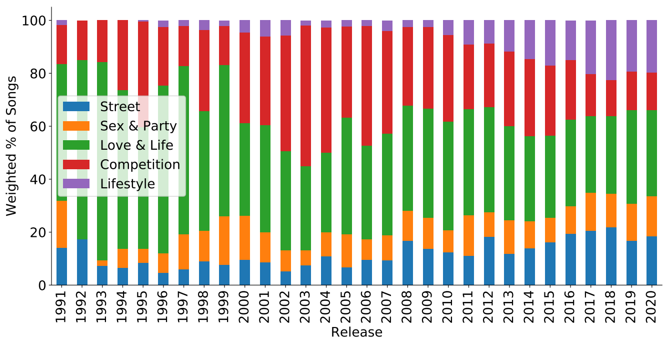

Fig. 2. Normalized timeline of the average topic distribution in the lyrics of songs released between 1991 and 2020. Songs are weighted by their popularity score on Spotify.

A. Topics in Rap over Time

In this section I visualize the development of German Hip Hop lyrics over the last 30 years. Figure 1 depicts the average topic distribution of all songs that were released in the respective year. Since the number of songs released differs each year, the data is normalized to better capture the bigger picture. Note that they years until 2003 contain less than 100 songs each, whereas each year after 2011 has more than 400 songs in the dataset. This imbalance is not due to a poorly select subset of artists, but rather due to the rapid growth in popularity of German Rap in the last 10 years [6]. Multiple trends can be seen in this graphic. First of all, an increase of the topic ”Competition” with a peak in 2005, followed by a steady decline until 2018. Next, we observe a fast growth of the ”Lifestyle” topic since 2010, as well as a moderate growth of ”Street” Rap between 2008 and 2018. Throughout most years, ”Love & Life” appears to be the most dominant topic.

This method automatically gives higher influence to artists with more songs in a release year. It also disregards the music preferences of the audience. Figure 2 is meant to capture the preferred topics by the consumers, by weighting each song with its Spotify popularity score. The score is downloaded for each song from the Spotify API using the *spotipy* package [7]. It measures the current popularity of a song on a scale from 0 to 100. To increase the influence of more popular songs on the graphic, I extended the scale to 1000. An exponential function maps e.g. the score 100 to 1000 or 90 to ca. 500. Compared to the unweighted graphic, we immediately notice a much higher share of the topics ”Street” and ”Sex & Party” throughout the entire timeline. The topics ”Love & Life” and ”Competition” on the other hand seem to be less popular. ”Lifestyle” is mostly unchanged by the weighting.

B. Sido’s Story - Underground to Mainstream

In this section I want to lay the focus on one particular Rapper, namely Sido. Although he started his music career in the late 90’s, his solo career did not start before the end of 2003, which is why the graph in Figure 3 starts in 2004. 2004 to 2007 we can see an upwards trend of the topics ”Competition” and ”Sex & Party”, contributing to almost 80% of the Rapper’s lyrics in 2007. The following years, there is a decline of the two topics, which are being contested by ”Love & Life”. 2019 ”Love & Life” makes up around 60% of the lyrics on its own. ”Street” appears to always play a role in his songs, hovering around 15% throughout the years. ”Lifestyle” has the lowest contribution to his music in each year, apart from a 20% peak in 2012.

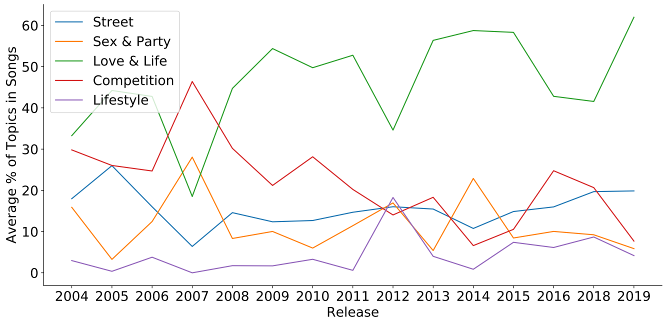

Fig. 3. The average percentage of topics in Sido’s songs between 2004 and 2019.

C. An Artist Overview

The last part of this section focusses on the similarity of the lyrics of all analyzed artists. Figure 4 visualizes the distance between the artists in 2D space. The average distribution of the artists’ songs for each of the five topics were the original vectors in 5D space, reduced to two dimensions using the transformation matrix obtained from Principal Component Analysis (PCA). The dots are colored with the majority label of each artist’s lyrics. We can see a large ”Love & Life” cluster on the left side of the graph, that makes up more than half of the entire space. The upper part is dominated by the ”Competition” topic, and the right side is shared by ”Street” and ”Lifestyle”. ”Sex & Party” as the majority label does not show up in an individual cluster, it only features two artists that are rather separated.

VI. Discussion

Analyse your results and discuss the possibilities and limitations of your technical approach. Compare your study to related work.

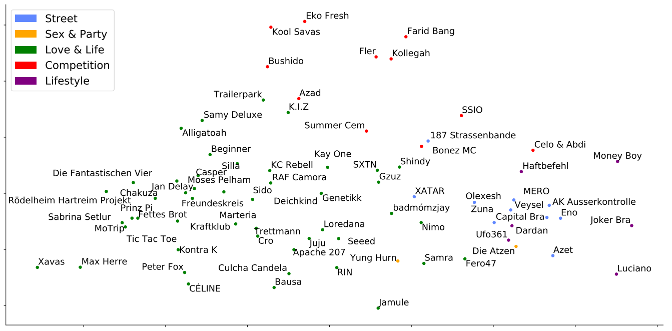

Fig. 4. A 2D visualization of the average topic distribution in each artist's lyrics. The dots are colored by the majority label.

A. *Analysis*

In this subsection I will give an interpretation to the previously shown results of section V.

1) Rap over Time: The rise of the ”Competition” topic in both Figures 1 and 2 strongly correlates with interesting developments in the German Rap-Scene, with the first rivalries that were carried out in public via Battle-Rap songs. The first big ”beef” was between Azad and Samy Deluxe in 2001. Azad provoked Samy Deluxe in his song ”Gegen den Strom” (against the stream), who then reacted with his disstrack ”Rache ist s” (revenge is sweet), which then again was countered by Azad with ”Samy De Bitch”. This is just one of countless battles that took place [8], with a peak after 2004, where famous Rapper Bushido left his label Aggro Berlin with help from the criminal Abou-Chaker clan [9].
”Lifestyle” is a topic mainly represented by artists of the Trap genre, which started to establish in Germany with artists like Money Boy, RIN, Yung Hurn and Ufo361. The growing popularity of this subgenre is also underlined by the fact, that each of the just named artists can be found in Spotify's ”Top Artists of 2020” playlist [2].
Streetrap has existed for a long time in Germany, as can be seen in 2, but has been through a revival that started with Haftbefehl in 2009. This revival phase was amplified by many now prominent Street Rappers like Bonez MC, Gzuz and RAF Camora [10]. Figure 1 also shows, that the volume of songs representing this style has only increased since then. From the popularity chart we can take away that the listening preferences of Spotify consumers show a strong bias towards the topics ”Street” and ”Sex & Party”. Not only recent songs, but also older ones seem to be more popular nowadays compared to contemporaneously produced songs.

2) Sido: The radical change in the lyrics depicted in Figure 3 correlates with Sidos development as an artist. Until the start of his solo career, he was wearing a chromium skull mask and known as a Gangsta-Rapper. His first album as a solo artist was furthermore heavily criticised and indexed [11] by the *Federal Review Board for Media Harmful to Minors* [12] because of misogynous and drug-glorifying texts. Today, the Rapper is regularly represented in the Charts with rather ”harmless” songs, often featuring Pop artists [13].

3) Overview:

VII. Conclusion

Based on your results and their analysis, what new knowledge do you take away from your project?

VIII. TODOBELOW

A. *Abbreviations and Acronyms*

B. *Units*

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. *Equations*

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. *LaTeX-Specific Advice*

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example" [3].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 5", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 5. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

References

Please number citations consecutively within brackets. The sentence punctuation follows the bracket.

References

[1] Wikipedia. (2020) Liste der meistverkauften Rapalben in Deutschland. [Online]. Available: https://de.wikipedia.org/wiki/Liste_der_meistverkauften_Rapalben_in_Deutschland

[2] Spotify. (2020) Top Künstler*innen 2020. [Online]. Available: https://open.spotify.com/playlist/37i9dQZF1DWTdV9tXbHOAv

[3] Genius docs. [Online]. Available: https://docs.genius.com/

[4] J. W. Miller. LyricsGenius: a Python client for the genius.com API. [Online]. Available: https://github.com/johnwmillr/LyricsGenius

[5] Github. (2018) Improve rule-based lemmatization and replace lookups. [Online]. Available: https://github.com/explosion/spaCy/issues/2668

[6] G. S. Dr. Florian Drcke, Sigrid Herrenbrck, “Musikindustrie in Zahlen 2019,” 2020.

[7] P. Lamere. Welcome to Spotipy! [Online]. Available: https://spotipy.readthedocs.io/en/2.16.1/

[8] Wikipedia. Liste von disstracks des deutschen hip-hops. [Online]. Available: https://de.wikipedia.org/wiki/Liste_von_Disstracks_des_deutschen_Hip-Hops

[9] J. Schaaf, “Bushido vs. arafat abou-chaker.” [Online]. Available: https://www.faz.net/aktuell/gesellschaft/kriminalitaet/bushido-vs-abou-chaker-ich-war-auf-dem-beifahrersitz-ganz-eingeschuechtert-16931630.html

[10] O. Marquart. Eine dekade im rückblick #3: Straenrap neue deutsche welle.

[11] NRWE. (2006) Verwaltungsgericht Kln, 27 K 6557/05. [Online]. Available: http://www.justiz.nrw.de/nrwe/ovgs/vg_koeln/j2006/27_K_6557_05urteil20060217.html

[12] “Tasks and responsibilities of the federal review board for media harmful to minors.” [Online]. Available: https://www.bundespruefstelle.de/bpjm/meta/en

[13] Billboard. Chart history sido. [Online]. Available: https://www.billboard.com/music/sido/chart-history/GES