The use of profanity in punk lyrics through years

Computational Literacy – Final Project

Jan Günther

1. Dataset and Research Question

For my research question, I want to know and answer the question of how profanity evolved through time in popular punk music. It is also interesting to me to see what kind of profanity is used over time and how this has evolved.

The dataset that I will use for this research question consists of two different acquired datasets. First, I will use the data provided by https://musicbrainz.org/, which is a not-for-profit collaborative music database. To access the data, there are two different ways to do that: the first would be accessing the data via the API. The second option I used in the end was to download the full data dump that is also provided by MusicBrainz. The full export is available at https://data.metabrainz.org/pub/musicbrainz/data/fullexport/. The data set I use here is limited to its search of punk songs to the years 1976 - 2017, as before and after this date, there is only a small number of songs that are tagged with "Punk" inside their database.

The second dataset that is used is the dataset that contains the lyrics data, which I want to analyze. The dataset I use here was provided by huggingface.co. It is described on the website as follows: "This dataset consists of roughly 480k English (classified using nltk language classifier) lyrics with some more metadata." It is available at https://huggingface.co/datasets/brunokreiner/genius-lyrics, and the data inside is sourced from genius.com.

Lastly, a source for profanity words was needed. For that, I found a list that included not just a list of words but also the category of that word, as well as its severity level. The information for that can be found here https://github.com/dsojevic/profanity-list/tree/main

2. Accessing and transforming the data

First approach to access the data

At first, I wanted to access the data from MusicBrainz via their API. The data obtained was then stored in a CSV file that contained all the information I needed. For my approach, I needed the following information: release year, title, artist, ranking, and number of votes. I needed this information so I could limit my analysis to the most popular songs of each year. This would also be important as the dataset from MusicBrainz has a large variety of songs that would not all be available in the lyrics dataset, and sorting for the most popular would remove non-relevant information for the search.

I switched the approach later and moved to downloading the full datadump for the MusicBrainz database since the limit rate for accessing information is 1 second per request, which led to a large time to access all the information I would need, as I would need one to two requests per entry to get all the information I needed. This whole procedure would take around two days for data from 1980 to 2021. I still include the

Python scripts for this approach in the repository if someone is interested in that.

Using the data dumps and other raw data

After discarding the first idea, I went on and loaded the data dumps into a database. For that, I used DuckDB as my DB system of choice, as it allows easy use inside a Python environment and does not require much setup. I imported the obtained datasets through Python scripts that used SQL commands and added views for those afterwards. I did the same for the CSV that included the lyrics of the songs, as well as the JSON list that included the information about the profane words.

When the data was imported, I could finally use it to start working on getting an answer to my question. I also needed to link the data set with lyrics to the data set for punk songs. For that, I used the provided ID in the lyrics set and added the corresponding ID to the table that included the recordings and their information. After that, I once again created views that would help me to get such information as "how many profane words were in each song," which would be done through a regex search that searches each lyric for words that are in the profane word list. I would also note down which categories of profanity would be included in the song, e.g., religious or sexual. The file also used a severity scale from 1 – 4, which I adapted so I could measure the average level of profanity.

At the last step, I used the Python library matplotlib to visualize the data to be able to see its trends and have a better visual understanding of the data I just transformed.

3. **Analyzing the data**

When we now look at the data and analyze it, we can view it from different perspectives. First, let's take a look at the average use of profane words throughout the year (Fig. 1). Here, we can see a small increase in profanity that took place around 1990, but at the same time, 1980 was the year with the highest average profanity in its lyrics. So, all I can say is that there was a low use of profanity between the years 1980 and 1990. After that time, the numbers show a rather normal change between years that I would consider as not noteworthy.

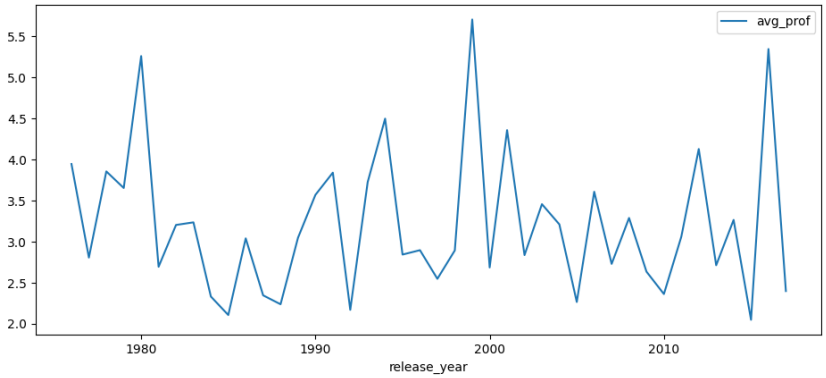

Figure 1 Average profanity use in lyrics over the years

To maybe also get a better understanding of the profanity increase over the years, it is also a good idea to look at the average density of profanity per word inside the lyrics (Fig. 2). This helps to see if there is actually an increase in profanity use or if the texts are just shorter/longer. While we can see a higher peak in 1990 compared to just the count of profane words, it shows in general that there is not really a blind spot in the other graph. It also once again shows the drop of profanity in the 1980s, even though it is not as clear as it was when just counting the number of words.

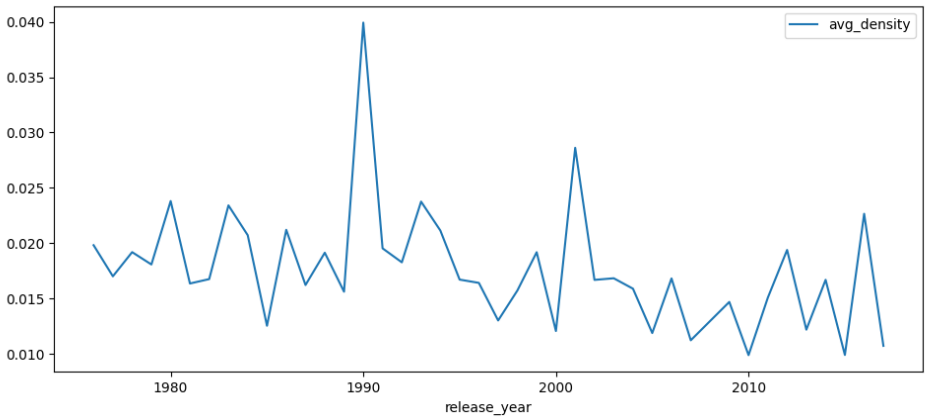

Figure 2 Average density of profanity per word over years

Secondly, when we take a look at the average severity of the profanity through the years (Fig. 3), we can see a more defined increase in the severity of the profanity when used starting after 1990. While there were some peaks before that, there are just a few dips in later years, with a more consistent use of more severe words after 1990. What can also be seen here is that while we have a general higher density in the year 1990(Fig. 2), the severity of those words shows a clear difference, with it being on the lower end of the scalar.

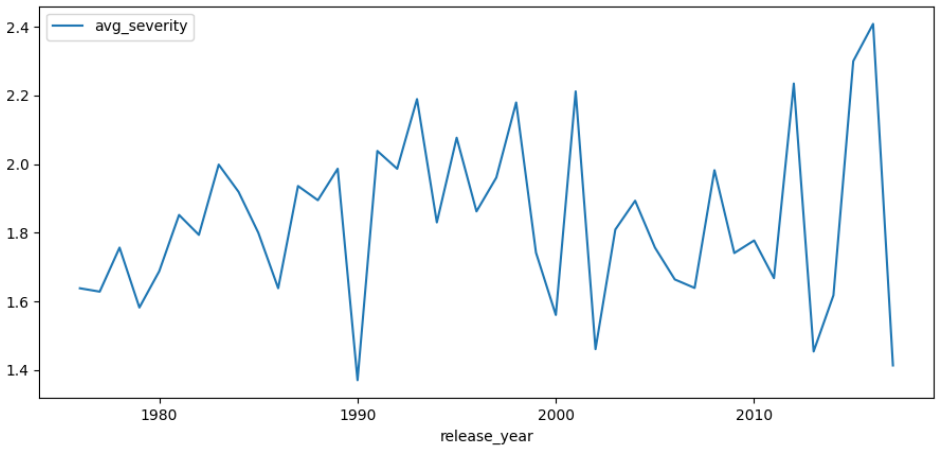

Figure 3 Average Severity of the profanity over the years

We can also take a look at the kind of profanity that is used in the lyrics, for which the list of profane words sets 5 different categories of origin (Fig. 4). These categories are “general”, “lgbtq”, “racial”, “religious”, and “sexual”. What can be seen here is that there is no huge shift where a category is changing over the years. We can also see that profanity that is typical against groups of people is, in general, not much included in the lyrics of popular songs in the category of punk.

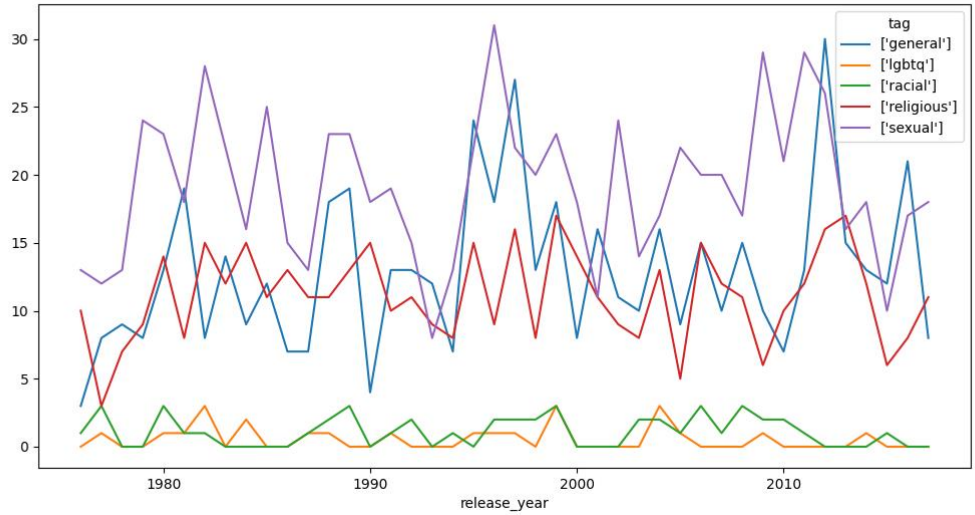

Figure 4 Appearance of profanity categorized over time

4. Discussion

When looking at the data, it seems that there is not much shift that is clear in one of the last few decades. But if anything, it helps to see that there is no huge shift in the last 30 – 40 years. In the 80s and 90s there was a moral panic of how modern music is to profane and the requests for censoring music was at a high, for example in a paper by Shoshana D. Samole it is argued that “The decade of the 1980's ushered in a new wave of shocking sexuality and aggression in popular music.“ (Samole, 1996, p. 176). When we take a look at the *obtained* data, this is a statement that does not appear to be reflected in comparison to popular punk music. Here we can see a clear spike in the year 1980 (Figure 1), but it normalizes after that and goes even below the values that were measured in lyrics from before those years. Monk-Turner and Sylvertooth (2008) also argue that „Sexual content in popular music has been existed since 1920s [...] music has the highest sexual contents compare to other media.“ (Monk-Turner & Sylvertooth, 2008, p.4) Which is something that is also observable when we look at Figure 4 where profanity that is tagged as sexual is the one with the most occurences through lyrics over time.

And while studies show that there is an increase in how often we use profanity in the English language (Deckker & Sumanasekara, 2025, pp. 8-11), it appears that there is, at least in popular Punk music, no increasing trend in the last 30 – 40 years when we look at the data.

5. **Data Limitation and improvements**

The data that was used for this project was a small sample size. This was due to the limitations on obtaining lyrics and music information in general. Most of the data is locked behind APIs and is not openly available; in addition to that, lyrics are also copyrighted material, so not all songs have freely available lyrics that can just be downloaded. This meant that I was limited to a dataset that was available to me. To have a clear understanding of the use of profanity, bigger sample sizes would be needed, so this can be seen as a test for the concept of analysing lyrics and using the MusicBrainz dataset as a source of music that is declared as "punk".

References

Samole, S. D. (1996). *Rock & roll control: Censoring music lyrics in the '90s*. University of Miami Entertainment & Sports Law Review, 13(2), 175–194. https://repository.law.miami.edu/umeslr/vol13/iss2/1

Monk-Turner, E., & Sylvertooth, D. (2008). Rap music: Gender difference in derogatory word use. *American Communication Journal, 10*(4), 1–12. https://digitalcommons.odu.edu/sociology_criminaljustice_fac_pubs/19

Deckker, D., & Sumanasekara, S. (2025). *Profanity through time: A corpus-based and sociolinguistic study of the evolution, usage, and perception of English curse words*. International Journal of Southern Economic Light, 13(5), 1–23. https://doi.org/10.36713/epra21959