Table of contents

Semantic and Sentiment Analysis of Song Lyrics using NLP

1. Introduction

The study of music lyrics has long served as an indicator of cultural, social, and emotional change. Songs are not only a source of entertainment, but they also encode social values, collective emotionality, and identity (DeWall, Pond, Campbell, & Twenge, 2011). Recently, large-scale lyric datasets and advancements in natural language processing have allowed researchers to move from experimental interpretations to systematic, data-driven analyses (Mihalcea & Strapparava, 2012; Fell & Sporleder, 2014). By analyzing the lyrics computationally, researchers can get insights into larger cultural phenomena that show how sentiment fluctuates, themes bundle, and social conflicts that are projected onto popular music develop over decades (Hu & Downie, 2010; Hunke, Heidemann, & Schreiber, 2025).

This study contributes to the growing development by employing topic modeling and sentiment classification on a compilation of English pop and rap lyrics from 1970 to 2020. It utilizes methods of Non-Negative Matrix Factorization (NMF) for semantic theme extraction and VADER sentiment analysis for sentiment polarity, and aims to empirically analyze how lyrical themes correspond to sentiment over time, if they do.

Music is personal and social. As an active listener of many genres, I am interested in how lyrics articulate ways emotions and values are framed. On a societal level, analyzing song lyrics is important because popular music has impacts on norms and perceptions, whether in building youth subcultures (“microcultures”), or normalizing beliefs about gender, religion, or violence (Christenson, de Haan-Rietdijk, Roberts, ter Bogt, & Meeus, 2012). Knowing how themes and sentiments have changed also provides a way to measure the currents of culture, and how they coincide with economic and social factors.

1.1 Research Questions and Relevance

RQ1. What are the main lexical-semantic themes, how they developed, and how they are connected to sentiment?

RQ2. How do these patterns link to broader social and cultural change?

This research aims to identify and label coherent lyrical themes using topic modeling, and to facilitate interpretability. Additionally, goal is to measure how these themes correlate with sentiment classes using coefficients (R^2 and p-values for robustness), and examine how the distribution of themes and their respective sentiments changed over time to point out cultural shifts (e.g. explicit language use in rap, romantic expression in pop).

Popular music is a unique cultural artifact and point of change when it comes to social attitudes; they contextualize dominant feelings, values, and identities across time. Studies show that lyrical content reflects wider social movements: for example, increases in selfish-ness and aggression within songs over the last few decades have been reported in America based on their local top charts (DeWall, Pond, Campbell, & Twenge, 2011). Content analyses of top-40 songs reveal that topics, such as love, rebellion, and materialism exhibit similar cyclical trends, thus reflecting societal issues and youth subcultures (Christenson, de Haan-Rietdijk, Roberts, ter Bogt, & Meeus, 2012). Therefore, song lyrics offer a unique and insightful factor for capturing current cultural mood and generational change.

Another indication of relevance addresses the economic significance of lyrics. The global music industry is a part of the “attention economy,” where products compete for scarce (and now incredibly short) consumer attention (Varian, 2020). Knowing which themes resonate with audiences can help determine marketing strategies, decisions for copyright, and decisions relating to the direction of streaming platforms.

From the perspective of methods, lyrics also offer a good trial ground for natural language processing (NLP) and data science. Opposed to lengthy documents, such as novels or political speeches, song lyrics are shorter, stylized, and repetitive, which poses different challenges in terms of text analysis (Hu & Downie, 2010).

Hence, the study’s significance lies at the intersection of cultural sociology and computational linguistics. Systematically analyzing lyrics across five decades, this project adds to cultural evolution and demonstrates how advanced approaches for NLP can help show patterns that simple reading may not uncover.

1.2 Note on theory

To illustrate how song lyrics can be analyzed systematically, it may be useful to outline the main conceptual and mathematical foundations of text representation, sentiment, and topic modeling.

In order to analyze any text computationally, it has to be converted to structured numerical representations of data. One approach is called the bag-of-words model, which represents each document as a matrix of word counts in the structure of a vector. The TF-IDF (term frequency-inverse document frequency) weighting scheme is used to eliminate frequent but not informative words:

$$tf - idf(t, d, D) = tf(t, d) \cdot log(\frac{N}{df(t,D)})$$

where $tf - idf(t, d, D)$ is the frequency of term t in document d, $df(t, D)$is the number of documents containing t, and N is the total number of documents (Manning, Raghavan, & Schütze, 2008). The document-term matrix produced serves as input for statistical modeling. Cosine similarity is a common way to calculate similarity between documents or between topics and words: $cos(\theta) = \frac{A \cdot B}{||A||B||}$, measuring is the angle between two vectors in high-dimensional space. This is a very common metric in NLP to assess relatedness of meaning.

Sentiment analysis is used to classify text according to the polarity of the text, typically positive, negative or neutral (Liu, 2012). Rule-based systems depend on lexicons of words that are associated with sentiment scores. The sentiment analysis algorithm used in this thesis is VADER (Valence Aware Dictionary and sEntiment Reasoner), which utilizes a sentiment lexicon supported by rules about how to treat negation, intensification, and punctuation (Hutto & Gilbert, 2014). For instance, positive or negative affects are expressed, i.e. “good” is positively polar and “not good” is negatively polar because of the negator.

Limitations must be acknowledged. Sentiment models can be weaker for irony, slang, or meanings that are contextual. Presenting categorical polarity in a positive, negative, or neutral fashion oversimplifies the emotion inherent in all lyricism. Lyrics often reflect more complex forms of experiences, such as longing, nostalgia, or spiritual connection (Dodds & Danforth, 2010). Nevertheless, sentiment analysis is particularly fitting to capture general orientations associated with long-term changes, which is the goal of this analysis.

Topic models seek to uncover latent topics (or themes) that account for the observed patterns in the distribution of those words. The aim is to approximate the document–term matrix $X \in R^{n\times m}$ with n documents and m terms as a product of two lower-rank matrices:

$X \approx WH$, where $W \in R^{n\times k}$ represent the document-topic matrix, $H \in R^{k\times m}$ denotes the topic-term matrix, and k is the number of topics.

Latent Dirichlet Allocation (LDA) treats the document as a mixture of topics and the topic itself as a polynomial distribution over words, which is estimated using Bayesian inference (Blei, Ng, & Jordan, 2003). Non-negative Matrix Factorization (NMF) on the other hand solves the optimization problem:

$$min_{W,H\geq 0} ||X - WH||_F^2$$

where $// \cdot //_F$ is the Frobenius norm, which ensures it is interpretable as both W and H are non-negative (Lee & Seung, 1999). NMF tends to produce more coherent topics for short, repetitive text, such as song lyrics (Fell & Sporleder, 2014).

Both sentiment analysis and topic models boil down to the application of classical statistics. Measures such as correlation coefficients or OLS regression are useful to determine if the relationships between topics and sentiment are reliable and not random.

2. Existing literature

Empirical work on song lyrics has grown dramatically in the last two decades in step with expanding large digital corpora and improvements in NLP. Studies have analyzed the lyrics in the context of broader cultural or psychological dynamics utilizing both qualitative interpretation and quantitative analysis.

Author DeWall, Pond, Campbell, and Twenge, (2011) conducted an analysis of the linguistic frequency results into Big 100 charts, which they reported growing self-focus and antisocial behavior as well as anger over 1980 to 2007 based on category lists. LIWC (Linguistic Inquiry and Word Count) is the method used in the study and works by counting frequencies of linguistics features of psychologically relevant categories (e.g. anger words). They applied regression models on LIWC scores for years, and determined significant upward trends of self-referential language and aggression.

Author Christenson, de Haan-Rietdijk, Roberts, ter Bogt, and Meeus, (2012) conducted content analysis using U.S. Top-40 songs from 1960 to 2010. They analyzed themes such as romance, rebellion, and materialism, and then applied proportion time series and chi-square to identify significant shifts in themes among songs. Their findings similarly contribute to cyclical shifts in songs and theorize that lyrics reflect existential patterns or music paradigms, suggesting a relationship with music and generations.

Sentiment lexicons have now been employed on the original lyrics (Dodds & Danforth, 2010). They utilized an approach which created a "hedonometer" (emotions scores) to evaluate happiness from text by applying word frequencies onto independent happiness scores. They have demonstrated how average happiness scores are decreasing across decades in popular music by utilizing weighted average scores and bootstrapped confidence intervals and have examined songs, blogs, and speeches.

Hunke et al. (2025) most recently used lexicon-based sentiment analysis and utilized word embeddings to analyze German song lyrics from 1954 to 2022. Their methodology combined dictionary-based sentiment assignment with distributional semantics (embedding scores) and word2vec embeddings to determine contextual similarity among sentiment words. They used linear mixed-effects models for statistical inference, explaining variability of repeated measures by artist and years.

The first computational methods to extract topics utilized Latent Dirichlet Allocation (LDA), which operates from estimating Dirichlet-distributed priors of each word-topic and document-topic distribution (Blei, Ng, & Jordan, 2003). For example, Fell and Sporleder (2014) applied LDA to lyrics in several languages and conducted clustering analysis of the corpus by mood and genre. They also used and evaluated several coherence metrics (pointwise mutual information (PMI) and perplexity) as reference points and found that topical modeling was able to explain stylistic traits that were not known solely from audio properties.

3. Methods

Song lyrics are treated as text data and analyzed using a modular, two-component pipeline that combines unsupervised topic modeling with lexicon-based sentiment analysis in order to examine how the emotional tone of pop and rap music in English-language songs has changed from 1970 through 2020, and how that tone moves together within lyrical themes. The design is straightforward:

- build a balanced and exposure-based corpus of the songs whose Genius pages were viewed the most each year (Top-100 per {genre, year})
- compute sentiment labels at the song level using VADER, a rule-based model that is calibrated to short, informal English and which is responsive to intensifiers, negation, and punctuation (Hutto & Gilbert, 2014);
- calculate a Non-negative Matrix Factorization (NMF) model across TF-IDF features
- connect the songs' topics to sentiment at two levels of granularity:

- (i) song level (correlations between topic weights and song sentiment labels)
- (ii) year level (regressing the average year sentiment distributions on the year's topic composition).

We maximize representativeness of mainstream exposure (the songs with the most Genius pageviews come closest to a measure of what the public consumed every year in music). The method supports comparability-to-metric across decades, by fixing the partner lyric/denominator-per-cell constant across decades. Third, the fixed metric keeps the TF-IDF × NMF problem so that in practice it would be possible to include different K's (amount of topics), and vocabulary thresholds with metrics under real life compute capabilities. Note that pop and rap are the two best represented genres in the original dataset and are particularly of my own interest as a consumer. We fixed K (number of topics) equal to 7 to ensure we have a well balanced overview.

3.1 Corpus construction and sampling

The corpus starts from the Genius Song Lyrics [song_lyrics.csv] with Language Information dataset from Kaggle, which ultimately collects Genius lyrics with meta data and pageview counts, as well as language information (CarlosGDCJ, 2022). We obtain only rows tagged pop or rap and with release years between 1970-2020, to capture the past 5 decades of music.

With respect to language, we restrict English in two stages. When both language identifiers are present, the lyric only survives if both CLD3 and fastText agree "en". If only one classifier is available, then "en" from that classifier is the only one needed. This conservative gate keeps the multi-language noise presentation low, and provides good alignment with VADER's English focus.

In order to maintain acceptable exposure each year and to maintain a consistent denominator, the corpus retains the Top 100 tracks, by {genre, year}, ranked by the Genius page views. Within each cell, with tokens (artist, song_name) normalized, we remove duplicates, resulting in around 9417 total songs.

English is selected to keep my measurability valued (most of the metrics of interest are English-based) and to keep from having issues of cross-linguistic comparability. Pop or rap are selected because they are the two most represented genres in the Kaggle/Genius dataset, and because the research questions I'm exploring center on the difference of their respective lyrical conventions. Top-100 per cell to maximize mainstream sample namespaces while managing TF-IDF vocabulary and NMF matrices, allowing enough room to maximize K and similarity measure robustness.

3.2 Text Normalization / Features (TF-IDF)

Because both lexical sentiment and topic models are sensitive to superficial variation, the lyrics are normalized prior to analysis. Square bracketed stage directions (e.g.; "[Chorus]"), parentheticals, and URLS are removed, unicode punctuation is normalized, and redundant whitespace is collapsed. Tokens will maintain inner apostrophes so that "don't" and AAVE forms like "ain't", "talkin'", "gettin'" will be able to be analyzed. A standard English stop list will be removed, with the exception of negators and polarity shifters (“no, not, never, without”). A small domain stop list only stops non-lexical fillers and ad-libs (“oh, la, skrrt,” etc.) that do not make much sense. WordNet lemmatization reduces inflectional variance, while preserving genre/dialect markers where appropriate.

Sentiment is analyzed with VADER: for each song, VADER will return positive, negative, neutral, and ambiguous proportions which will sum to 1, and a compound score $c \in [-1, 1]$. The label mapping of the compound score is: positive if $c \geq 0.20$, negative if $c \leq -0.20$, neutral if $|c| < 0.05$, ambiguous otherwise (to account for mixed-valence texts where both the positive and negative sub-scores are non-negligible while the total compound is nearly zero). While lexicon-based approaches cannot reliably capture sarcasm (e.g., irony or parody) or genre-specific pragmatic use, VADER for negation scope, boosters, capitalization, and repeated punctuation show competitively good performance on short informal English, which is just right for song lyrics lines (Hutto & Gilbert, 2014; Pang & Lee, 2008).

Recall that each song d is represented as a unigram TF-IDF vector over a pruned vocabulary in the topic modeling subsection. Terms were filtered with min_df = 10 to remove very rare forms and max_df = 0.7 to remove very common items. To filter residual non-English noise, we only kept tokens whose wordfreq Zipf scores exceeded length-dependent threshold. Stopwords (except for negators) were filtered again in the vectorizer.

3.3. Topic modeling (NMF)

Hyperparameters were the number of topics K = 7, a NNDSVD initialization, a coordinate-descent solver, and 600 iterations. The rows of W were normalized to L1, so that the topic weights for each song sum to one for interpretable proportions. For topic descriptors, the top large values of H were used after a minimal local-specificity filtering (to downweigh very short or common strings). Human readable labels were the high-salience unigrams.

3.4 Topic-model-sentiment procedure

We connect lexical themes (topics) to sentiment, implemented in the two analysis scripts. Inputs are the row-normalized NMF topic weights for each song (i.e., columns topic_0…topic_6) and a four-class VADER label (positive, negative, neutral, and ambiguous) assigned from the normalized lyric text, using the thresholds.

The first script generates a document-level topic by sentiment association table. For each topic k and each sentiment class s, it treats the song’s proportion of topics as the continuous dependent variable $y_d = W_{dk}$, and the sentiment assignment as a binary variable $x_d^{(s)} = \mathrm{I}\{sentiment = s\}$. Two complementary statistics are calculated: Pearson’s $r_{k.s} = corr(y, x^{(s)})$ with the two-tailed p-value; and the slope β_1 from the OLS model $y_d = \beta_0 + \beta_1 x_d^{(s)} + \varepsilon_d$, reported along with model R^2 and the two-tailed p-value for β_1. Pearson’s r captures the linear association intuitively, while β_1 is the difference between the mean value of topic weight of songs in class s, and all others.

The second script generates within-topic sentiment time series. For each topic k, year y, and sentiment l, the topic-weighted share is constructed as follows:

$$WithinTopicShare_{k,l,y} = \frac{\Sigma_{d \in D_y} W_{dk} \, \mathrm{I}\{sentiment = s\}}{\Sigma_{d \in D_y} W_{dk}}$$

which estimates the proportion, among material that are closely related to k in year y that are labeled l. By weighting with W_{dk}, songs that are more “about” a topic tend to be weighted more in that topic’s yearly sentiment series.

These two methods offer, respectively, a cross-sectional snapshot of how topic focus co-varies with sentiment labels at the document level, and a long-term view into the framing of each topic’s changes over time.

4. Results

4.1 Sentiment Panels

Throughout the entire time frame, pop lyrics have consistently shown a high positive share, likely in the range of 0.65-0.75, and have shown only slight variations from a slowly decreasing long-run mean. The patterns between rap and pop are completely different: after a relatively high and volatile set of positive shares in the early 1970s, the rap series drifts from the late 1970s to the 1990s, and is quite a bit under pop after. By the 2000s-2010s, rap’s positive share is usually at or near the 0.30-0.45 range, supplemented by a slight uptick by the late 2000s, but there are no evident signs of potential future convergence of the two. [Appendix 1]

The opposite is true for that of negative sentiment. Simply put, rap shows a steady rise in the negative share from the late 1970s (exceeding 0.50 in the 1990s and around 0.50-0.70 after), and pop remains low and stable, around 0.20-0.35 range, and exhibits a slight upwards trend from late 1990s on. [Appendix 2]

Neutral and ambiguous categories are not recorded for either genre and are rare, usually 0–3% of yearly counts. Rare in the overall user-level sense of being captured, it means that movements of polarity (positive - negative) are only movements between the dominant being categories, no reclassification into neutral/ambiguous. [Appendix 3 & 4]

Decade graphics encapsulate the preceding patterns. For pop, the average positive share is stable. For the 1970s–2000s, the average positive share is about 0.70-0.75, slightly dipping in the 2010s, and average negative share is about 0.20-0.30 and trending slightly up into the 2010s. For rap, the decade averages exhibit the secular covariance from positive to negative sentiments, with the positive averages dropping from around 0.60 in the 1970s to around 0.35-0.40 by the 2010s/2020s, with negative averages raising from around 0.30 to around 0.60–0.70 on the same timeline. [Appendix 5 & 6]

To estimate long-run change, we estimate linear trends of year sentiment shares on centered decades (year/10, mean-centered), separately by genre via an interaction for centered decades. The fitted trend lines are overlaid on the observed data, and the slope coefficients are interpreted as the change in percentage points over decades.

The estimated slope for pop is a -1.29 percentage points / decade (pp/dec), that is a slight long-run decrease from a higher starting level. The slope for rap is a -6.49 pp/dec slope which matches the pronounced downtrend listed above. The between-genre difference in positive slope therefore is -5.20 pp/dec for rap compared to pop. [Appendix 7]

For pop, the negative share grows at +1.55 pp/dec modestly. For rap, we see a slope of +6.59 pp/dec; similar to the raw series, it suggests there is a large secular increase present in rap. The implied rap-pop differential is +5.03 pp/dec. [Appendix 8]

The polarity trend sustains the previous two panels: pop polarity declines at -2.85 pp/dec, while rap polarity declines at -13.08 pp/dec. Thus, rap polarity becomes negative over the sample, while pop is positive but declining. The estimated rap-pop differential in polarity slope is -10.23 pp/dec. [Appendix 9]

4.2 Residual diagnostics

Scatterplots of OLS residuals against year provide a non-parametric structure check not accounted for by the linear specification. Regarding the positive share, pop residuals range narrowly around zero with only low-amplitude systematic variation. In contrast, rap residuals reveal significant positive deviations in the late-1970s to early-1980s (actuals above linear fit) and negative deviations in the early-mid-1990s (actuals below linear fit). Rap residuals moved closer to zero in the 2010s. This indicates that a single linear trend is a parsimony summary but smooths over sub-period curvature in rap.

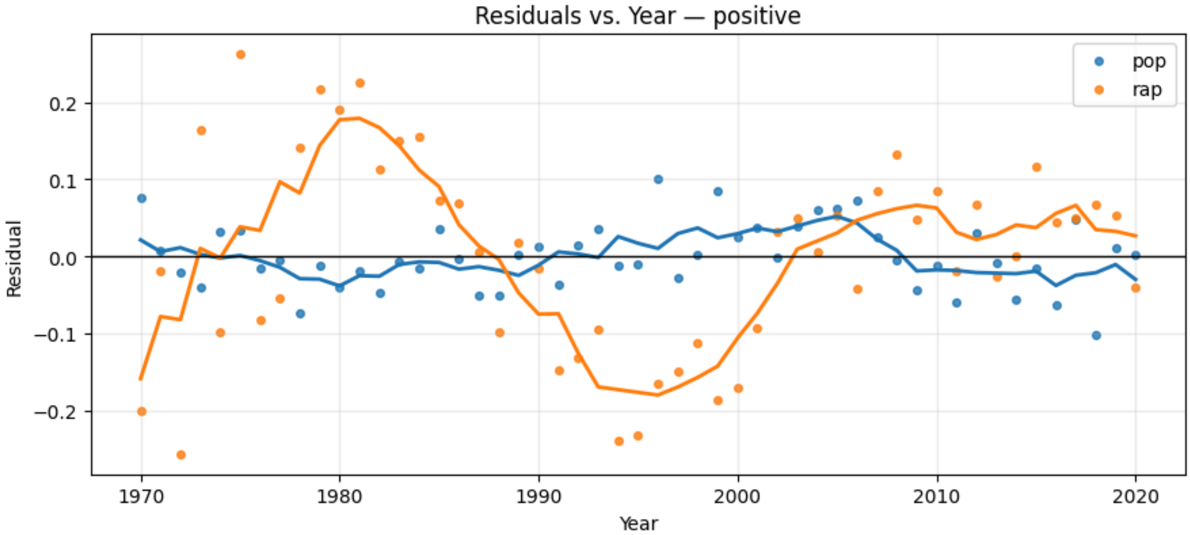


Figure 1. Negative sentiment OLS residuals over years for 2 genres

For the negative share, pop again shows small residual structure; rap shows positive residual clusters in the late 1980s-1990s and negative residual clusters in the late

1970s and late 2000s.

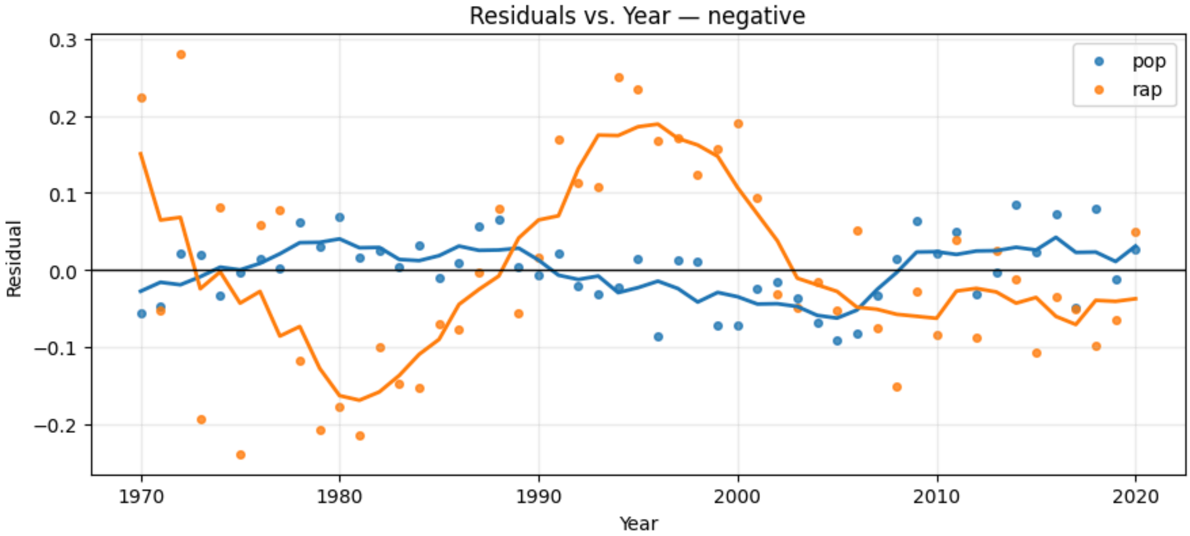


Figure 2. Negative sentiment OLS residuals over years for 2 genres

The polarity residuals shed light on those patterns: pop residuals are nearly zero with a slight dip in the last portion; rap residuals have a U-pattern: positive in the late 1970s/early 1980s, negative in the 1990s-early 2000s, and returning to near 0 in the 2010s again suggesting sub-period dynamics in addition to a regular line.

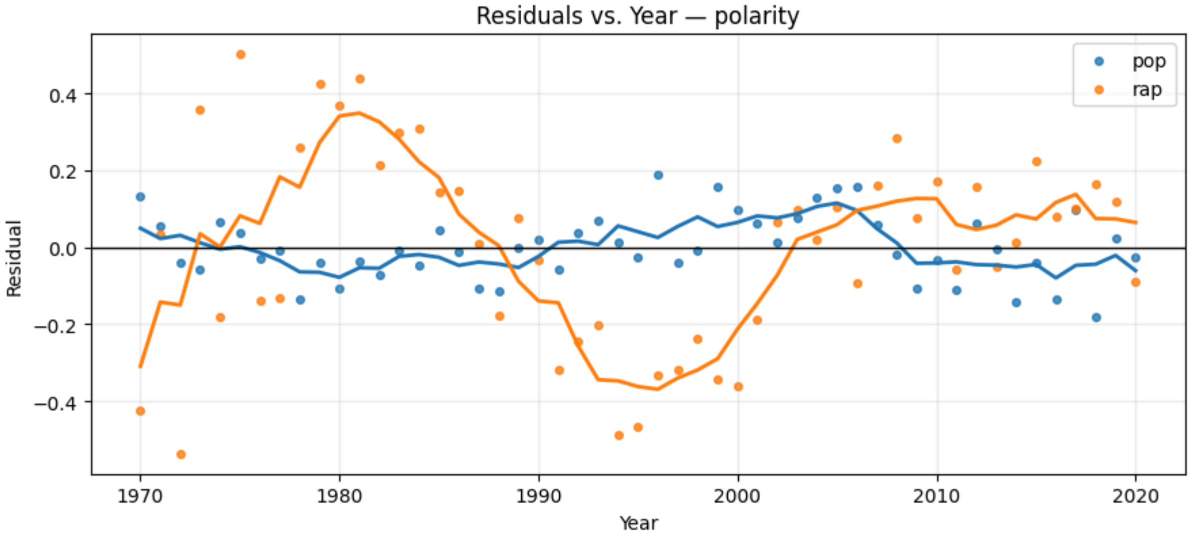


Figure 3. Polarity (positive-negative) OLS residuals over years for 2 genres

Together, the residual panels validate the direction and strength of the genre trend, but also demonstrate rap is proven non-linear (especially 1978–1999) which a linear specification brushes over. The implications are for future work, segmented/stepwise regressions or splines, but these modeling decisions will not be covered in this section.

With no neutral/ambiguous shares doubling count, the genre gap in positivity is really just reducing the genre gap in negativity by absolute number and differences in polarity. Three descriptive regularities come up:

- The level gap is stable: in recent years, pop is more positive than rap and rap is less negative than pop, with polarity gaps in the 0.40-0.70 range most commonly after 1999.
- The gap in trend is large: OLS slopes imply that the long-run trend decline in rap positivity and growth in rap negativity is about 5 pp/dec stronger (in absolute value) than pop.
- Variability is greater in rap, particularly prior to 1990, recounted by the greater swings residuals in Figures 1-3.

4.3 Topic dynamics (TF-IDF + NMF)

Topics are constructed by factoring the TF-IDF term-document matrix with non-negative matrix factorization (NMF) and result in a topics-by-term matrix H (top terms per topic) and a documents-by-topics matrix W (topic weights per song). For ease of reading, each topic is titled by the 3 highest weight terms from H; bar plots show the top-k terms and top H_{kj} weights; yearly “topic share” curves show the average W weight per topic within year; stacked area charts show the yearly composition (shares sum to 1); and a heatmap illustrates topic intensity across years. **Disclaimer: sensitive language**. Resources include abusive or explicit words. In the thesis text, they are masked (e.g. “b*tch”) for reader comfort, but the analysis uses the original tokens to measure accurately. The thesis covers popular lyrics, so references to profanity, sexualization, and insulting slurs are quoted. These terms are repeated as is, for the sake of accuracy in linguistic measures. Their inclusion is only reporting what is in the corpus, not the author's views.

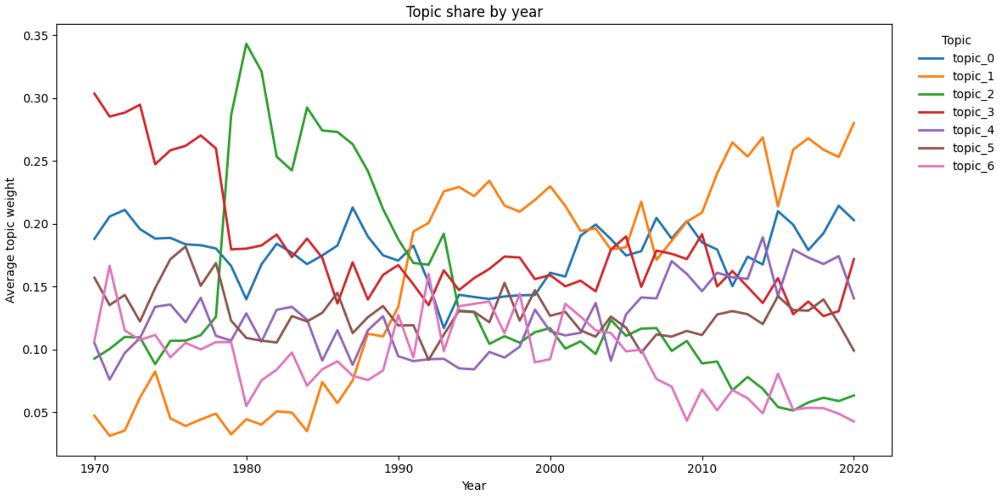


Figure 4. Topic share by year across 7 themes

4.4 Topic inventory (Top Terms)

There are 7 stable lexical themes identified:

Topic 0 - “Think / Sorry / Thing”. High-weight terms: “think, sorry, thing, wrong, friend, not, forget, never, try, understand”. This collection of terms reflects interpersonal and intrapersonal processes and attempts to process relational regrets. [Appendix 10a]

Topic 1 - “B*tch / N*gga / P*ssy”. Terms appear vulgar (“b*tch, n*gga, p*ssy, f*ck, sh*t, m*therfucker, d*ck, ain’t, shoot, money, g*ng, smoke”), indexing an explicit/violent experience. [Appendix 10b]

Topic 2 - “Funky / Party / Disco.” Dance/party vocabulary (“funky, rock, party, disco, rhythm, rhyme, beat, boogie, dance, groove, mic, master, house”), seems to be related to dance and parties. [Appendix 10c]

Topic 3 - “Dream / Light / Night.” Lyrical, image-studded vernacular (“dream, sun, light, sky, night, morning, rain, world, moon, day, wind, shadow”), connects to metaphorical and atmosphere writing, yet is the most generalized of all. [Appendix 10d]

Topic 4 - “Tonight / Wanna / Right.” Conversational pop (“baby, tonight, wanna, babe, girl, right, want, gonna, gotta, love, kiss”) presents flirtation and romance. [Appendix 10e]

Topic 5 - “Heart / Apart / Forever.” Heartbreak ballad vocabulary (“heart, love, apart, forever, feel, lonely, break, hurt, tears, promise, true”). [Appendix 10f]

Topic 6 - “Praise / Jesus / Worship.” Religious vocabulary (“praise, jesus, lord, worship, holy, glory, bless, christ, god, grace, heaven, spirit”). [Appendix 10g]

4.5 Yearly topic shares, Heatmap

Line plots of average topic weight by year signal shifts in lexical focal points over the years 1970-2020 (Figure 4). Three trends stand out:

Increase in explicit/violent slang (Topic 1). The topic is low in the 1970s, takes off around the early 1990s, and stays floating during the 2000s-2010s. By the end of the time-series, Topic 1 is one of the highest-share topics, and exhibited peaks in intensity in the 2010s.

Decrease of party/dance vocabulary (Topic 2). Topic 2 peaks around the late 1970s-early 1980s, in accordance with the disco/dance period, and then sharply declines over the 1990s and 2000s; stabilizes at low levels after that.

Relative consistency of quotable pop and reflection (Topics 4 and 0). Topic 4 (“Tonight/Wanna/Right”) trends positively around the 1990s, and contributes a significant share beyond the year 2000; Topic 0 (“Think/Sorry/Thing”) is steadily present with minor variation and no radical changes.

Two additional trends are identifiable at medium levels: Topic 3 (“Dream/Light/Night”) show gradual declines with higher amounts in the 1970s-1980s; Topic 6 (“Praise/Jesus/Worship”) trends softer over years, with higher intensity pre-1990 and softer trend after. Topic 5 (“Heart/Apart/Forever”) is stable or trends positively, indicating a persistent but less frequent vocabulary of romance/break-up.

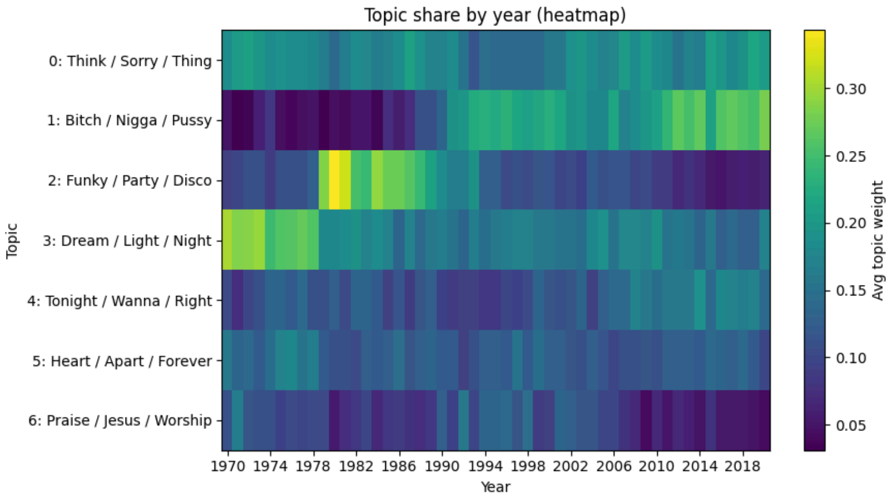


Figure 5. Heatmap of topic share by year for 7 topics

The heatmap, which ticks topic labels on the y-axis and year on the x-axis, helps us localize peaks. Topic 2 (“Funky/Party/Disco”) shows very high intensity during late 1970s to early 1980s (bright band), then is progressively getting darker afterwards. Topic 1 (“B*tch/N*gga/P*ssy”) has low intensity before 1990 and brightens intensely from the 1990s onwards, with the highest intensity in the 2010s. Topic 4 (“Tonight/Wanna/Right”) increases in warmth from the mid-1990s onward, and remains relatively warm through to the 2010s. Topic 3 (“Dream/Light/Night”) shows warm colorations in the 1970s-1980s and then fades afterwards. Topic 6 (“Praise/Jesus/Worship”) increases in warmth early and reduces later, showing a long-run softening of religious language. Topic 0 (“Think / Sorry / Thing”) was consistently mid-toned through the years, showing neither peaks or drops. Topic 5 (“Heart / Apart / Forever”) was moderately warm during early years with a slow decline after the 2000s.

4.6 Sentiment–topic association and within-topic polarity

4.6.1 Correlations between topics and sentiment shares (by yearly averages)

To summarize the ways that lexical types relate with sentiment over time, we correlate yearly average topic shares with yearly sentiment shares (VADER; positive/negative/neutral/ambiguous) from 1970-2020. Coefficients are Pearson’s r; 1,2,3 asterisks represent two-tailed significance ($p < .05$, $p < .01$, $p < .001$ respectively). “Positive r” and “negative r” discussed here refer to correlation to positive and negative sentiment respectively. [Appendix 11]

First, the explicit/violent slang topic, “B*tch / N*gga / P*ssy,” correlates negatively with positive sentiment and positively with negative sentiment (Positive r = −.430***; Negative r = +.450***). The effect sizes are the largest absolute values of all topics, which implies that when this lexicon has greater prominence within a year, the corpus’ index of negative shares tend to be higher and positive shares are lower.

Second, lexicons surrounding imagery, “Dream / Light / Night,” disco/dance, “Funky / Party / Disco,” romance/break-up, “Heart / Apart / Forever,” religious register, “Praise / Jesus / Worship,” and conversational pop, “Tonight / Wanna / Right,” all showed to be positive correlations with positive sentiment and negative correlations with negative sentiment. Effect sizes vary meaningfully: the “Heart / Apart / Forever” category has the largest positive alignment (Positive r = +.188***; Negative r = −.189***), this is followed by “Tonight / Wanna / Right” (Positive r = +.122***; Negative r = −.121***) and “Dream / Light / Night” (Positive r = +.120***; Negative r = −.137***). “Funky / Party / Disco” and “Praise / Jesus / Worship” have smaller, but directionally stable coefficients (all significant at p < .001).

The third observation is the reflective/apologetic category, “Think / Sorry / Thing,” with show the opposite sign; mildly negative with positive sentiment and mildly positive with negative sentiment (Positive r = −.042***; Negative r = +.036***). While the magnitudes were small, the signs indicated that on average every year, a greater presence of this topic had a marginally increased negative share. Most years neutral and ambiguous shares are near-zero and typically weakly, mostly insignificant with regards to correlations with topics (the few significant cells are trivially small, e.g., ambiguous with “Dream / Light / Night” at r = +.042***; “Think / Sorry / Thing” at r = +.024*). Hence, the two most significant cases of sentiment-topic coincidences are (i) negative alignment with the explicit/violent slang topic as well as (ii) positive alignment with romance, image, and conversational pop lexicons.

4.6.2 Changes in sentiment composition by topic over time

Next, we take a closer look at the sentiment composition by topic - that is, for each topic and year, the weighted share of positive/negative/neutral/ambiguous sentiment among songs, using each song’s topic weight as the within-topic weight. This also indicates whether or not the polarity of a topic changes from decade to decade, regardless of overall prevalence.

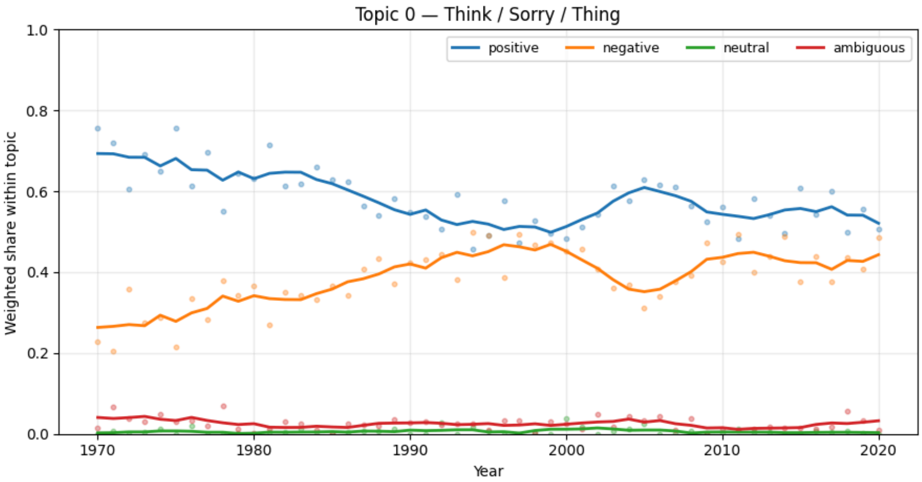


Figure 6. Changes in sentiment in topic 0

Topic 0 - “Think / Sorry / Thing”: Positive dominates the sentiment, but it has a drop around the mid-1990s (positive goes down toward ~0.50 and negative rises to ~0.40-0.45). Positive rebounds nicely in the mid-2000s to bring us back to ~0.60; then it slides a little to the late 2010s. Neutral and ambiguous stay near zero.

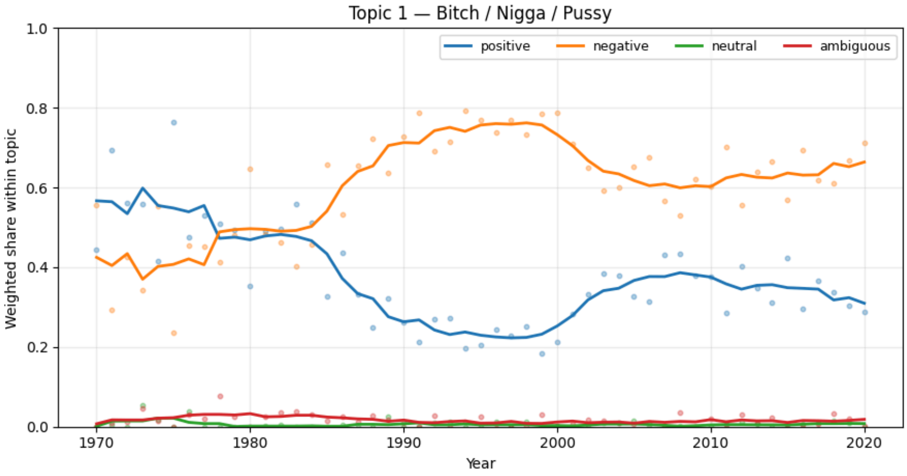


Figure 7. Changes in sentiment in topic 1

Topic 1 - “B*tch / N*gga / P*ssy”: This one has a structural polarity flip. Positive falls from the 1980s-1990s, down to ~0.20-0.30, while negative is hair rising to ~0.70-0.75 in the 1990s and continues to remain dominant (>0.60). The within-topic follows the correlation results: when this lexicon gets triggered, the internal sentiment composition is persistently negative.

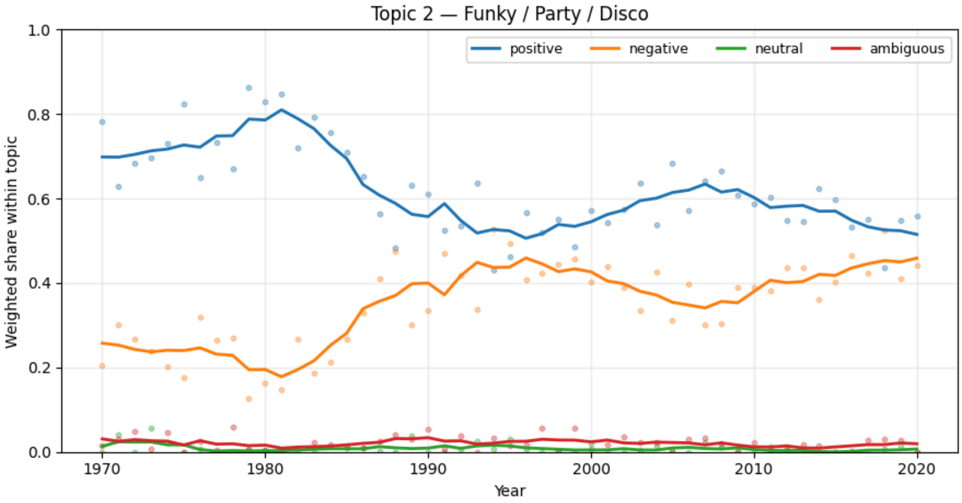


Figure 8. Changes in sentiment in topic 2

Topic 2 - “Funky / Party / Disco”: Early decades are very positive (often ≥ 0.70-0.80) and very little negative. Then around the late 1980s-early 1990s, positive recedes to ~0.50-0.60 with negative gains. A partial recovery in the mid-2000s (positive ~0.60-0.65), then it slipped back down again in the late 2010s. Neutral and ambiguous have almost no scores.

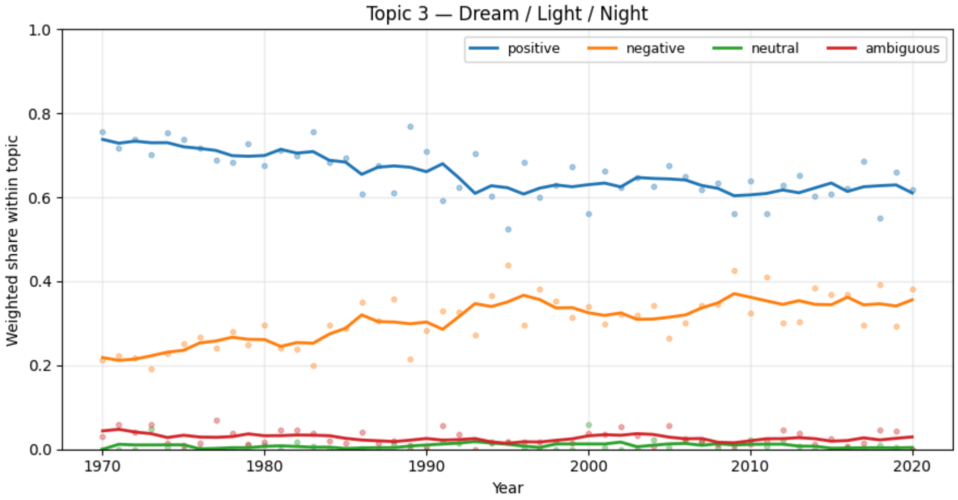


Figure 9. Changes in sentiment in topic 3

Topic 3 - “Dream / Light / Night”: The sentiment is within bumpers and positive dominates. Positive shifts downward gradually from the mid-0.70s to low-0.60s by the 2010s. Negative added from ~0.20 to ~0.35-0.40. Positive is the dominant within-topic sentiment all years despite the convergence.

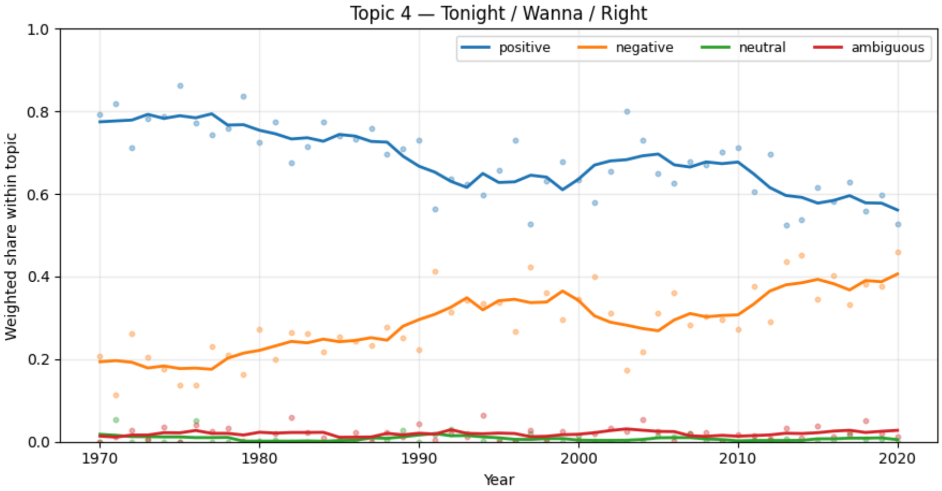


Figure 10. Changes in sentiment in topic 4

Topic 4 - “Tonight / Wanna / Right”: Another positive dominant topic, which falls: positive shifts from ~0.80 in the 1970s to ~0.55-0.60 around ~2020. Negative shifts from ~0.20 to ~0.35-0.40. The trajectory suggests an internal slow darkening.

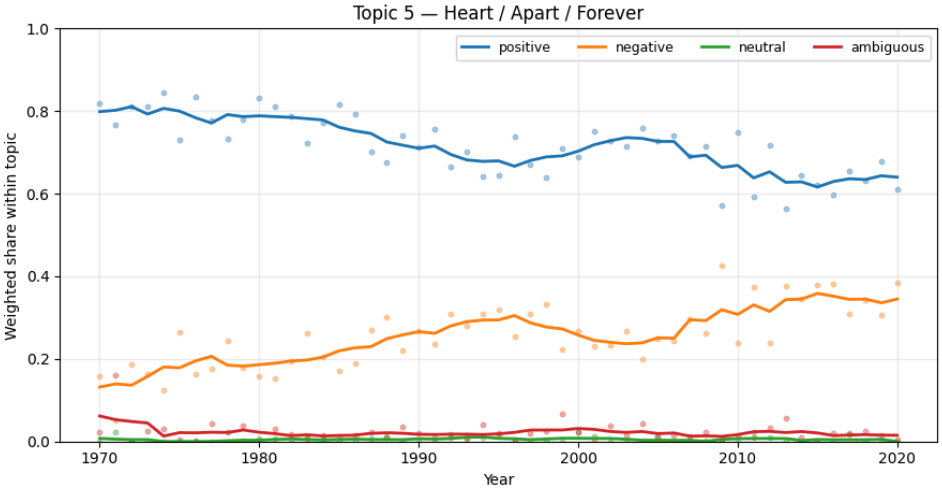


Figure 11. Changes in sentiment in topic 5

Topic 5 - “Heart / Apart / Forever”: This was a consistently positive topic in the early sample (positive ~0.80+), falling slowly to ~0.65 by the 2010s, while negative moves from ~0.15 to ~0.30-0.35. Positive is still the dominant within-topic class all years.

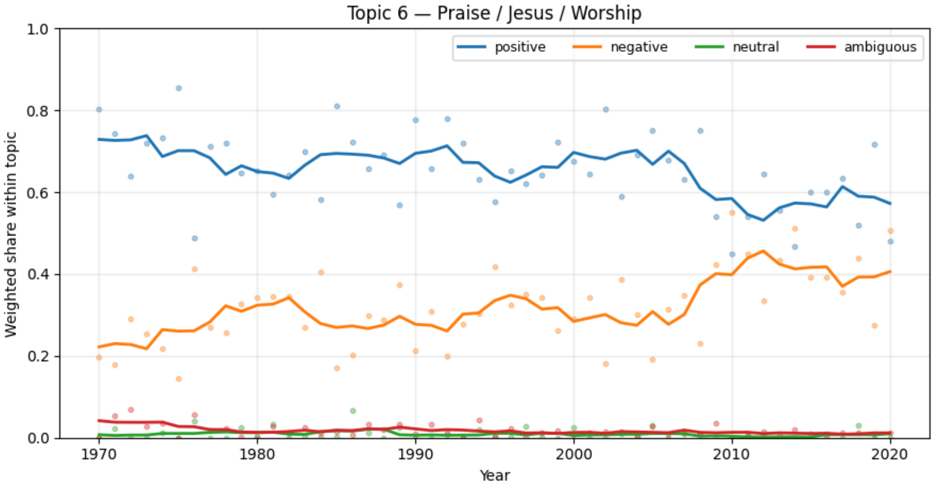


Figure 12. Changes in sentiment in topic 6

Topic 6 - “Praise / Jesus / Worship”: This trend starts positive (0.70-0.80) and trends lower to ~0.55-0.60 by the late 2010s; and negative climbs from ~0.20 to ~0.35-0.40. The former parallels the long-term decrease in positivity of the religious topic.

Aside from Topic 1, all within-topic sentiments were positive-majority. However, most moved towards higher negative percentages since the 1990s, which fits the earlier presented corpus-level patterns of movement in earlier sections. The degree of movement to higher negative is variable, with Topics 4-6 experiencing the greatest shift to higher negative, and Topic 3 the least amount of shift to higher negative. Neutral and ambiguous lines for all topics and years are near zero and do not drive the same effects described.

5. Discussion

This chapter reinterprets the empirical results after aligning the seven NMF topics with the labels. For convenience purposes, RQ1 will be split into 3 sub-questions to address. From this point forwards, topics 0-7 will be called by terms most closely generalizing their respective top words:

- T0 - Self-Reflection (“Think / Sorry / Thing”)
- T1 - Violence/Profanity (“B*tch / N*gga / P*ssy”)
- T2 - Party/Dance (“Funky / Party / Disco”)
- T3 - Miscellaneous (“Dream / Light / Night”) - treated as the residual topic
- T4 - Flirtation/Love (“Tonight / Wanna / Right”)
- T5 - Heartbreak/Ballad (“Heart / Apart / Forever”)
  T6 - Religion (““Praise / Jesus / Worship”)

5.1 RQ1: Predominant Topics (RQ1a–c)

What are the predominant lexical-semantic topics, how did they develop, and how do they relate to sentiment?

RQ1a. What are the predominant lexical-semantic topics?

The NMF decomposition finds seven interpretable topics that are also stable “parts” of lyrics as non-negative factorization (Lee and Seung, 1999, 2001). Taking the highest-weight words in the term-topic matrix HHH, we find the topics:

“Self-reflection” represents a language of thinking back and potentially indulging in human interactions (apologizing, discussing, confronting someone). It is a general relational register that spans across artists and decades and manifests in verses where blame, overthinking or self-doubts are worked through.

“Violence/profanity” features explicit language, slurs, and some Afro Anglo Latino/ African American Vernacular English (AAL/AAVE) following slang (“b*tch, sh*t, ain’t, shoot, g*ng”). Semantically, it indexes confrontation and aggressive posture. Linguistically, it indexes in-group coded from rap and similar styles (Rose 1994; Kubrin 2005).

“Party/Dance” combines late-1970s dance/disco lexicon and early MC party talk (“funky, party, disco, rhythm, beat, boogie, dance, groove, mic, master”). The lexicon is unique in terms of density of rhythmic/performance nouns and is used in celebration contexts.

“Miscellaneous” aggregates imagistic and natural-world words (“dream, light, night, sun, sky, rain, world, moon, day:), the most “poetic” register of the corpus applied for mood and metaphoric effect. It is general enough, so can be considered the “catch-all” category.

“Flirtation/Love” encodes communicational intent within a romantic context (“baby, tonight, wanna, gonna, gotta, want, love, kiss”). It’s the carcass for many pop choruses and hooks.

“Heartbreak/Ballad” is the common break-up theme (“heart, love, apart, forever, feel, lonely, break, hurt, tears, promise”). It conveys both euphoria and grief/misery, but is lexically centered on relational status.

“Religion” gives the devotional register (“praise, Jesus, Lord, worship, holy, bless, glory, God, heaven, spirit”), found in crossover pop or in music made by religious artists of any genre.

Together, these seven topics encompass reflective talk, confrontationalism, dance culture, lyricism, romancing, relationship problems, and worship. Their durability across parameter initializations and their clear top terms demonstrate the model is not picking up artefacts but noticeable semantic “pieces” of popular music lyrics.

RQ1b. How did topic frequencies develop from 1970 to 2020?

Averaged yearly document-topic weights W, the stacked composition, and the heatmap all demonstrate reallocation of lexical attention over the decades. Overall, there are three movements.

1) A late-1970s peak and decline of dance/disco lexicon. “Party/Dance” ascends continuously, peaking around 1979-1981, before descending in the 1980s and stabilizes at near ~0.05-0.10 onwards in 2000. The shape of this arc mirrors the historical movement of disco - rapid mainstream propagation and a push-back at the boundary of the 1980s, and then dance idioms remain but do not dominate (Lawrence 2003; Shuker 2017). This topic also conveys strong positive affect and its outflow eliminates a considerable pool of upbeat language from the corpus, and it is at the beginning of the 1980s.

2) A late-1980s rise, and long plateau of explicit/insult slang. “Violence/Profanity” is low in the 1970s (around 0.03-0.08), rises at the end of the 1980s, and takes a step-increase beginning in the early 1990s, remaining high for the 2000s-2010s. The heatmap illustrates consistent high intensity for the final decades. The timing was coincident with the mainstreaming of hip-hop aesthetics in US charts around 1991, often credited as the stylistic “revolution” in quantitative histories (Mauch et al 2015).

A slow shifting of the lyric “middle” toward colloquial address and away from devotion and imagery. “Religion” turned downwards over time, with shares in the 1970s and single digits by the 2010s. In contrast, “Flirtation/Love” went upwards after ~1995, arriving at a steady state at ~0.15-0.20 layer by the 2000s. “Miscellaneous” which is relatively stable, declined from ~0.25-0.30 shares in the late 1970s to about ~0.08-0.12 shares by the 2010s. “Self-reflection” and “Heartbreak/Ballad” were relatively stable layers (typical layer share ~0.12-0.20) that showed modest drifts in share but no structural breaks.

Cross-year topic weights are normalized within a year, such that topic increases always reflect a displacement of other topics. Therefore, the stacked composition reveals that, as explicit slang and conversational hooks increased after 1990, disco, devotional and imagistic vocab extruded each other. This is important for affect because the topics dropping out were assigning positive associated valent, while increasing explicit slang were assigning negative affect.

RQ1c. How are topics connected to sentiment?

We address connecting topics to sentiment leveraging two ways.

First, annually correlating topic shares to sentiment shares produces a plausible matrix. The explicit/insult shares show the strongest associations of all the entries in the entire table, negative with positive and positive with negative (Positive r ≈ -.430; Negative r ≈ +.450). In contrast, “Heartbreak” is the only theme to show a strong positivity alignment (Positive r ≈ +.188; Negative r ≈ -.189), which is followed by “Miscellaneous” (≈ +.120/-.137), “Flirtation/Love” (≈ +.122/-.121), “Dance/Party” (≈ +.077/-.079), and religious (≈ +.083/-.082). “Self-reflection” is only weakly negatively skewed (≈ -.042/+ .036). Due to the rarity of neutral or ambiguous categories in the yearly aggregates, polarity change is effectively determined solely by the positive-negative balance, where the strength of the associations is the greatest.

Further, we take within-topic sentiment shares and aggregate those by timer (songs weighted by topic weights). Six of the seven topics are positive-majority within each year, but most have exhibited gradual internal darkening since the 1990s. In “Party/Dance”, “Flirtation/Love”, "Heartbreak", and “Religion” the positive line typically drops by 10-20 share points from the 1970s reference points, alongside an increase in negative. “Miscellaneous” drifts very little. The exception is the “Profanity/Violence” topic, which is negative-majority from the early 1990s on (often 0.60 to 0.80 negative).

We can now make logical inferences, tying sentiment and themes over the years. First, more years of explicit slang and less disco/religion are years with lower positivity and higher negativity indexes. Second, even if the same theme is dominant (love, lust, devotion), the way it is expressed shifts to the increased negativity. The corpus changes both what it talks about and how it talks about it.

Two notes related to methodology place bounds on magnitudes without changing conclusions. VADER performs well on informal English (Hutto & Gilbert, 2014), but like many lexicons, it may be overly taxable toward AAL/AAVE references and rechristened slurs (Davidson, Bhattacharya, & Weber, 2019; Sap, Card, Gabriel, Choi, & Jurafsky, 2019). This probably raises the negativity associated with the explicit/insult topic and rap more generally. Also, our NMF was trained on the pooled corpus; while timing strongly points to rap as the site to “Violence/Profanity” growth and popularity, the notebook does not derive genre-conditional topic proportions, thus genre should be viewed as an inference supported by the sentiment/chronological record, rather than as measurements of content.

5.2 RQ2: Links to Social & Cultural Change

How do these patterns link to broader social and cultural change?

The chronology and form of the lexical-sentiment changes correspond with well-recorded changes in the U.S./Anglophone popular music, and, society.

Quantitative musicology marks a stylistic “revolution” around 1991 in Hot-100 charts, with the mainstream rise of hip-hop aesthetics (Mauch, MacCallum, Levy, & Leroi, 2015). Industry reports show that R&B/hip-hop became the most-consumed genre in the U.S. by late 2010s; caused, in part, by the transition to on-demand streaming and playlist agreement (Nielsen, 2018; IFPI, 2019). Our data provides another instance of this change: "Violence/Profanity" spiked in the early 1990s, and the economic and statistical size of rap's negative trend is massive compared to pop.

Content analysis and cultural history shows how street realism, structural violence, and transgressive language are central to commercially bringing up the 1990s rap (Rose, 1994; Kubrin, 2005). Those aesthetic choices directly relate to the lexicon recovered in this topic and partly explains both the between-topic (greater T1 portion) and within-topic (majority negative) findings presented here.

Lexicon-based sentiment measurement tools including VADER can severely under- or miscalculate AAL/AAVE indicators and slurs, resulting in an exaggerated level of negativity for hip-hop styles (Davidson, Bhattacharya, & Weber, 2019; Sap et al., 2019). Timing and direction of our trends are aligned with outside musicology, but the effect sizes for T1 and the rap series should be exercised cautiously, given the demonstrated NLP bias in toxicity/sentiment analysis (Hutto & Gilbert, 2014; Kiritchenko & Mohammad, 2018; Sap et al., 2019).

The peak of T2 "Party/Dance" and subsequent decline transcript the typical history of disco: quick popularization in the mainstream in the mid-to-late 1970s and a public backlash in 1979-1980 (when "Disco Demolition Night" became a marker of cultural history) followed by incorporation into the large categories of dance/pop (Lawrence, 2003; Shuker, 2017). Given that this topic is very positively aligned as well as the majority lexicon is positivity-aligned, the depletion allows removal of a large set of happy tokens from the larger lexicon, serving as an explanation for the decline in positivity for the corpus through the softening of the upper body in the early 1980s.

T6 "Religion" allows to see a gradual decline from the 1970s until about the 2010s. To that end, this evidence is similar to what is found with declining Christian affiliation and the rise of religiously unaffiliated in U.S. survey data, notably from 2009-2019 (Pew Research Center, 2019). The positivity alignment, along with a drift toward slightly increased negativity, provides evidence that overtly devotional language is sprouting from charts in the mainstream, and that any remaining religious references are being reframed in slight ways.

The parental advisory program (mid-1980s onward) formalized explicit content labeling (RIAA, n.d.). In the broadcast era, the label sometimes limited airplay, while in the digital/streaming era, it largely functions as a content marker with modest distribution friction, thereby normalizing explicit strategies as stylistic markers, rather than barriers (Nielsen, 2018; IFPI, 2019). This institutional context is consonant with the extended elevation of T1 "Violence/Profanity" from the 1990s to 2010s illustrated on our heatmap.

Large-scale analysis of chart music identifies declining valence/"happiness" and increasing "sadness" through the mid-1980s to 2010s based on audio features and lyric proxies (Interiano et al., 2018). Our within-topic curves (T4-T6) illustrate just this internal darkening within otherwise "light" topics (desire, love), suggesting that stylistic reframing, not just topic substitutions, contributed to the long-run affect changes.

5.3 Limitations and implications

First, lexicon-based sentiment can misunderstand stance, and important attribution (e.g., irony, dialect) can sometimes mislead; the rather pronounced bias in race-dialect can encode bias in racial-dialect; therefore, we emphasize direction and timing more than absolute levels and report robust trend diagnostics. Second, the corpus emphasizes mainstream English releases; different models may not follow similar paths.

The results indicate that the emotional profile of mainstream lyrics follows both to who is absorbing first (the topic composition in response to genre change) and how the timeless themes are vocalized (the leveled-up reframing within topic). The tempo of mainstream hip hop, the post-disco pin, secularization, and content/platform regimes explain the origins and scale of change we observe (Lawrence, 2003; Mauch et al., 2015; Pew Research Center, 2019; IFPI, 2019). Methodologically, future work should include lexicon and dialect-aware context models (fine-tuned transformers), benchmark against human-coded lyric subsamples, develop a process to decode lyrical sentiment and musical affect (Interiano et al., 2018).

6. Conclusion

This thesis began with the goals of: (i) identifying coherent lexical-semantic strands in English language pop and rap lyrics from 1970-2020 and tracking their development (ii) connecting those strands to sentiment, at the corpus and within-topic levels. By using TF-IDF with 7-topic NMF and VADER sentiment labels, the project provides a fully transparent documentation of changes in lyrical content and sentiment over 5 decades.

This paper uncovered 7 consistent lyrical themes in pop and rap music in English (1970-2020), while also relating those themes to sentiment. On a broad scope, topic composition changed rather radically: “Violence/Profanity” increased significantly starting in the early ‘90s, whereas Party/Dance decayed after the disco era, “Religion” calmed down, “Flirtation/Love increased”, and “Self-Reflection” and “Heartbreak” remained stable. Years that had more “Violence/Profanity” were significantly less positive and more negative than other years; whereas years with more “Heartbreak/Love/Miscellaneous/Religion” and “Party/Dance” had more positivity.

And beyond composition style, the vast majority of topics darkened internally (i.e., down 10-20pp on positive share since the 1970s), so even if composition changed, how it was framed in the text changed as well. Pop was still broadly positive with slight softening. Rap style trended more sharply negative post-1970s.

These trends are harmonical with cultural trends (e.g. hip-hop becoming mainstream, Disco origin and backlash). Overall limits are lexicon biases (for instance, against AAL/AAVE), limitation of lyrical text to only English, and modeling decisions (K, pre-processing). Still, using a testable pipeline (TF-IDF + NMF + VADER) leads to verified insights. Future studies can expand using dialect aware, lyric meaningful sentiment, dynamic/genre conditional topics, and human-coded reference to get magnitudes while maintaining interpretability.

Ethics, Copyright, and Data Use

This dissertation uses publicly available song lyrics as text data and does not involve human subjects. Full song lyrics are not shared; to provide scholarly commentary, only brief, necessary quotations of song lyrics are shown. For measurement fidelity, the computational pipeline captures the tokens verbatim, though in written form, sensitive or derogatory terms are quoted (or masked). No raw song lyrics or bulk transcriptions are distributed; only code, derived features and summary statistics will be released. Data is kept locally, with limited access, and only retained as long as is required to complete analysis.

Bibliography

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://www.jmlr.org/papers/v3/blei03a.html

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. Foundations and Trends in Information Retrieval, 11(2–3), 143–296. https://doi.org/10.1561/1500000030

CarlosGDCJ. (2022). Genius Song Lyrics with Language Information [Data set]. Kaggle. https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information

Christenson, P. G., de Haan-Rietdijk, S., Roberts, D. F., ter Bogt, T. F. M., & Meeus, W. H. J. (2012). What has America been singing about? Trends in themes in the U.S. top-40 songs: 1960–2010. Psychology of Music, 40(5), 471–490. https://doi.org/10.1177/0305735612440612

DeWall, C. N., Pond, R. S., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. Psychology of Aesthetics, Creativity, and the Arts, 5(3), 200–207. https://doi.org/10.1037/a0023195

Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness Studies, 11(4), 441–456. https://doi.org/10.1007/s10902-009-9150-9

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1–26.

Fell, M., & Sporleder, C. (2014). Lyrics-based analysis and classification of music. In Proceedings of COLING 2014: Technical Papers (pp. 620–631). Association for Computational Linguistics. https://aclanthology.org/C14-1059/

Gillis, N. (2017). Introduction to nonnegative matrix factorization. SIAM Review, 59(4), 779–836. https://doi.org/10.1137/16M1080171

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.

Hu, X., & Downie, J. S. (2010). When lyrics outperform audio for music mood classification: A feature analysis. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010) (pp. 619–624). https://ismir2010.ismir.net/proceedings/ismir2010-100.pdf

Hunke, T., Huber, F., & Steffens, J. (2025). The evolution of song lyrics: An NLP-based analysis of popular music in Germany from 1954 to 2022. Music & Science, 8, 1–20. https://doi.org/10.1177/20592043251331155

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of ICWSM-14 (pp. 216–225). AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Kubrin, C. E. (2005). Gangstas, thugs, and hustlas: Identity and the code of the street in rap music. Social Problems, 52(3), 360–378. https://doi.org/10.1525/sp.2005.52.3.360

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791. https://doi.org/10.1038/44565

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13 (pp. 556–562). MIT Press.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mihalcea, R., & Strapparava, C. (2012). Lyrics, music, and emotions. In Proceedings of EMNLP-CoNLL 2012 (pp. 590–599). Association for Computational Linguistics. https://aclanthology.org/D12-1054/

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703–708.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

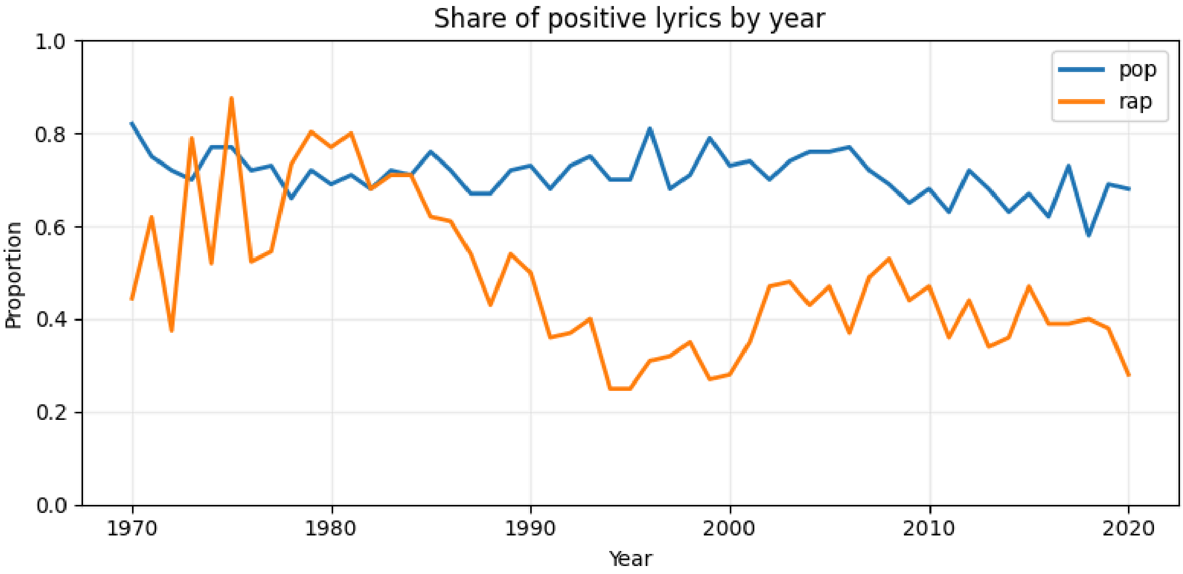
Pew Research Center. (2019, October 17). In U.S., decline of Christianity continues at rapid pace. https://www.pewresearch.org/religion/2019/10/17/in-u-s-decline-of-christianity-continues-at-rapid-pace/

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

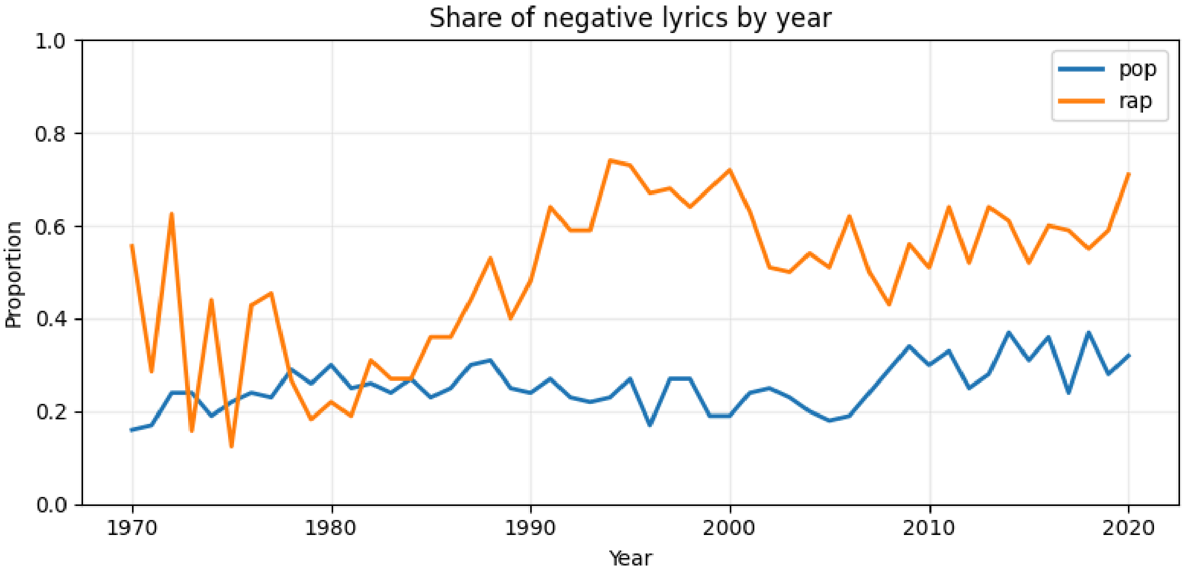
Speer, R., Chin, J., & Lin, C. (2018). wordfreq: Zipf-scaled word frequencies [Computer software]. https://github.com/rspeer/wordfreq

Varian, H. R. (2020). Intermediate microeconomics: A modern approach (9th ed.). W. W. Norton & Company.

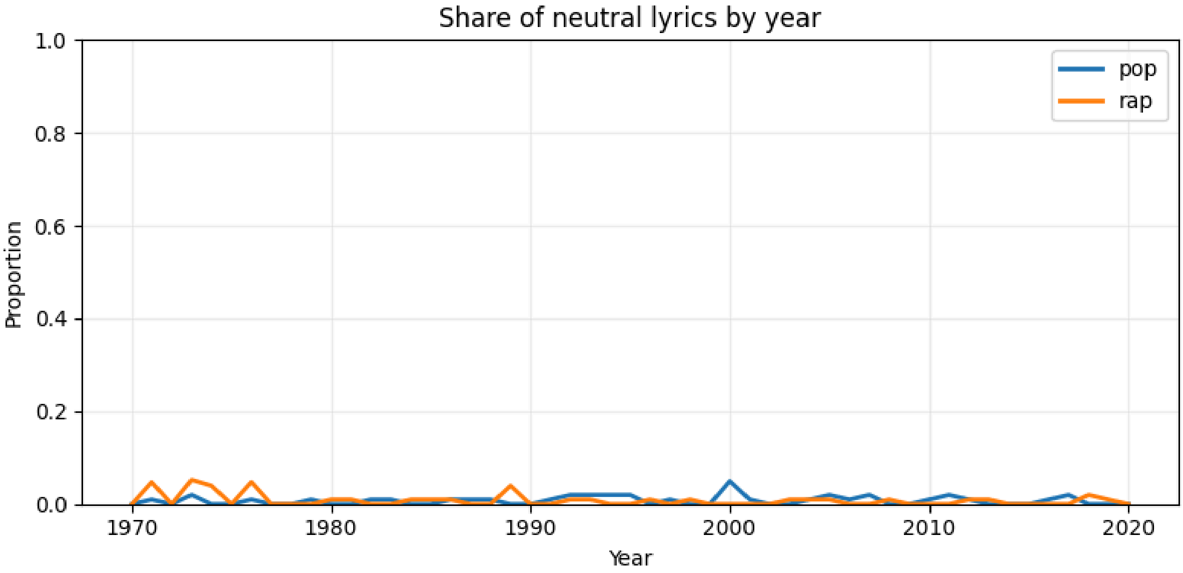
Appendices



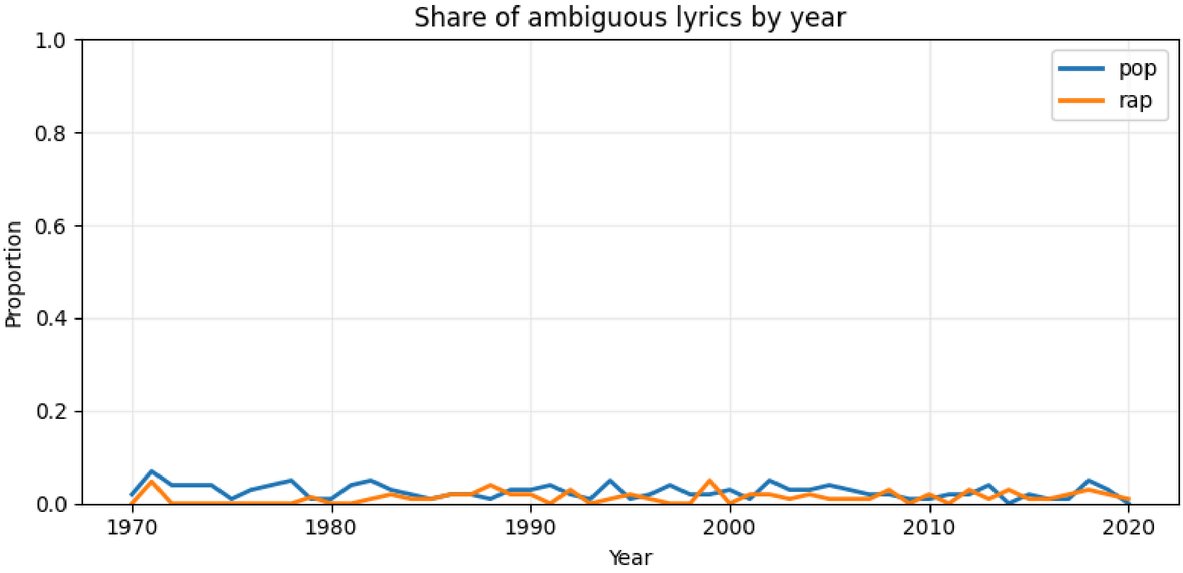
Appendix 1. Share of positive sentiment in lyrics by year for 2 genres



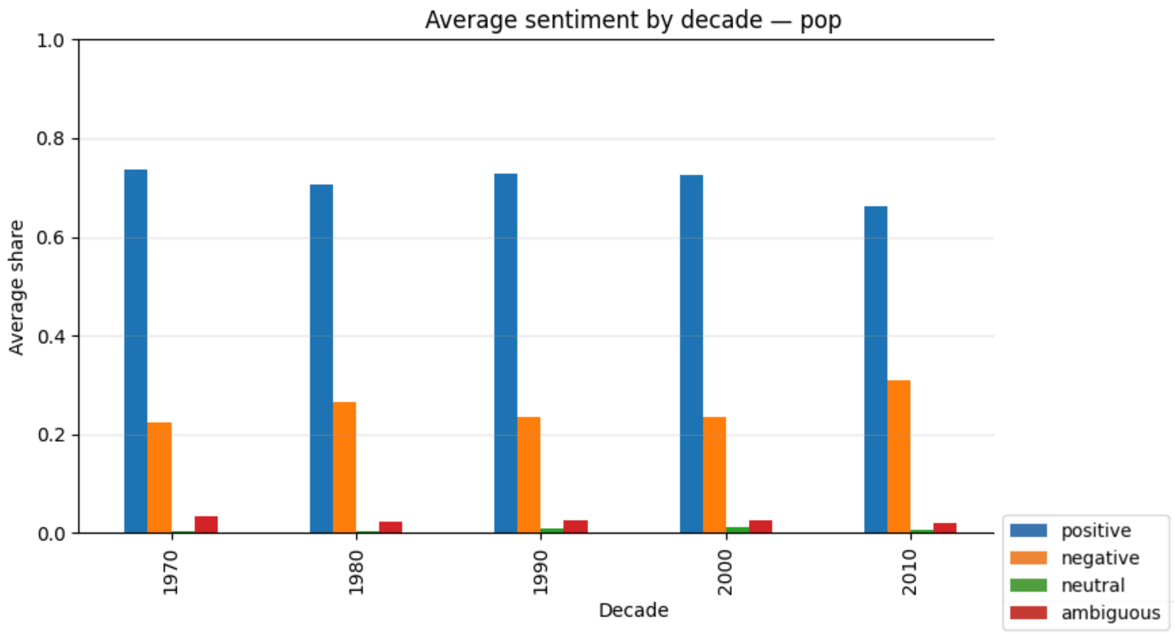
Appendix 2. Share of negative sentiment in lyrics by year for 2 genres



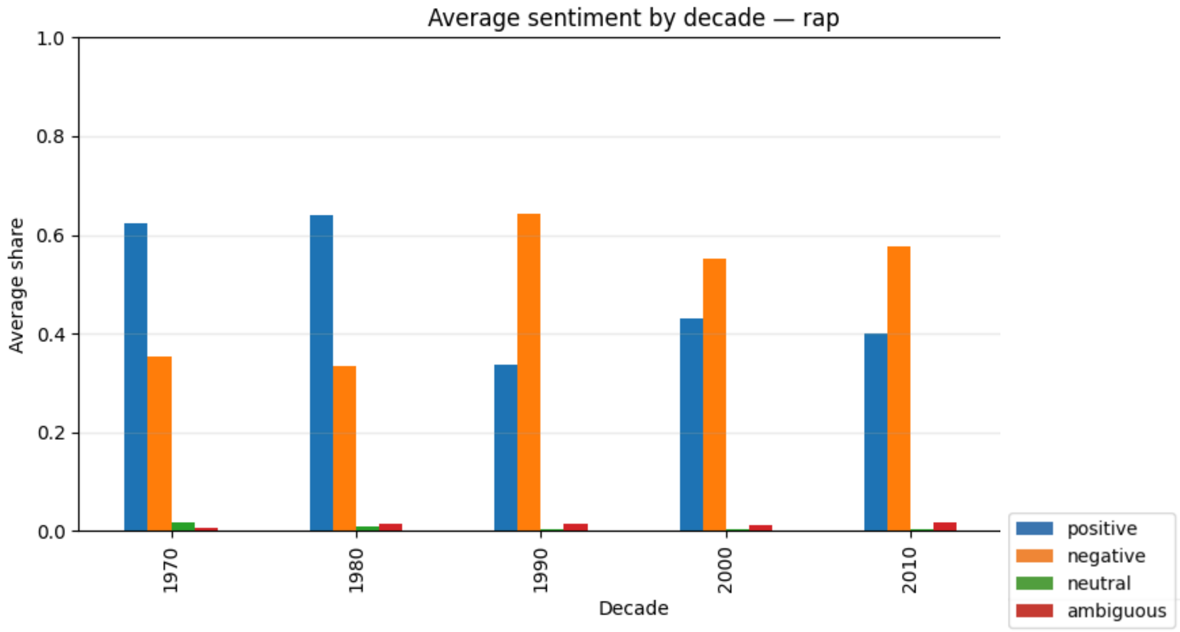
Appendix 3. Share of neutral sentiment lyrics by year for 2 genres



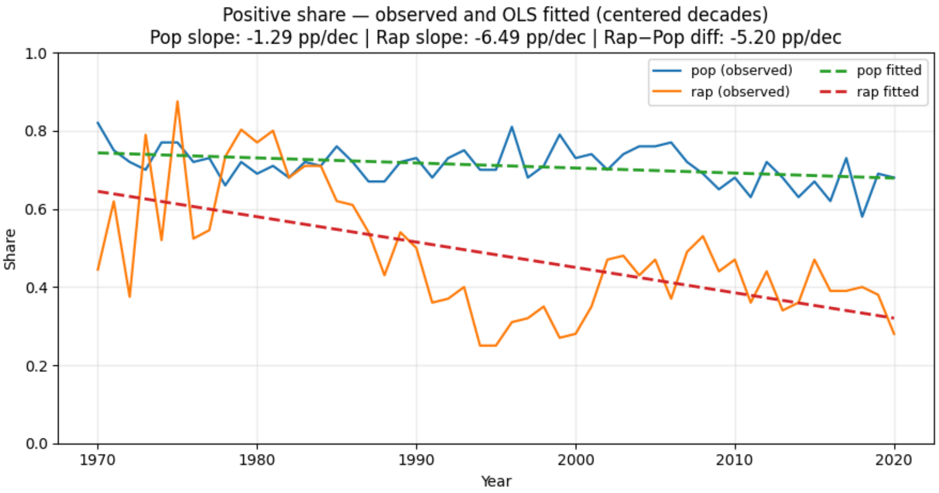
Appendix 4. Share of ambiguous sentiment lyrics by year for 2 genres



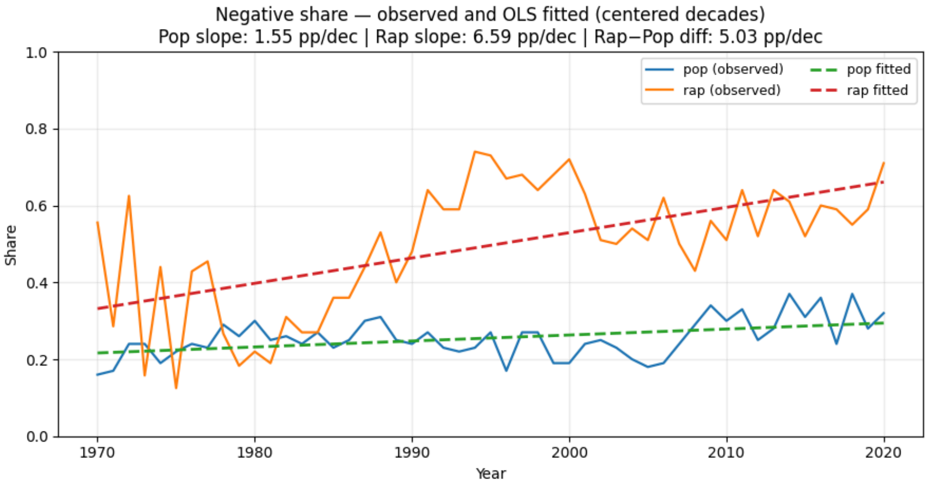
Appendix 5. Average sentiment by decade for pop lyrics



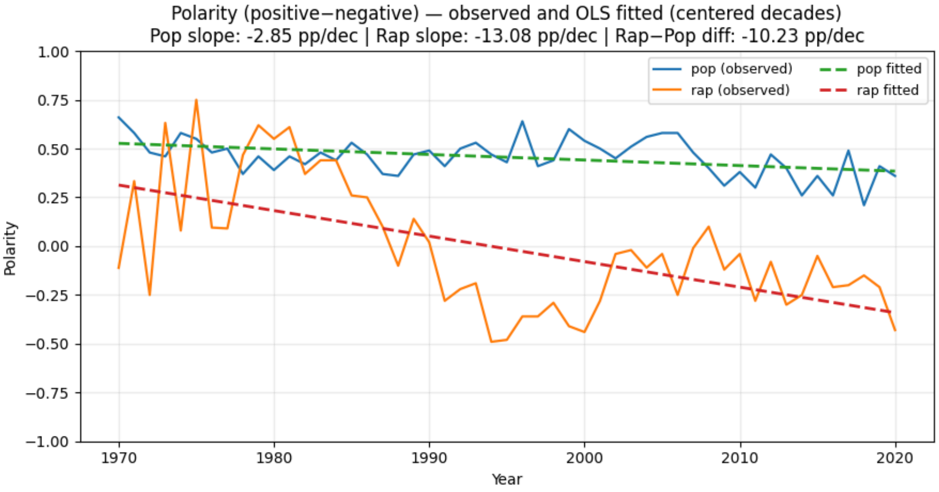
Appendix 6. Average sentiment by decade for rap lyrics



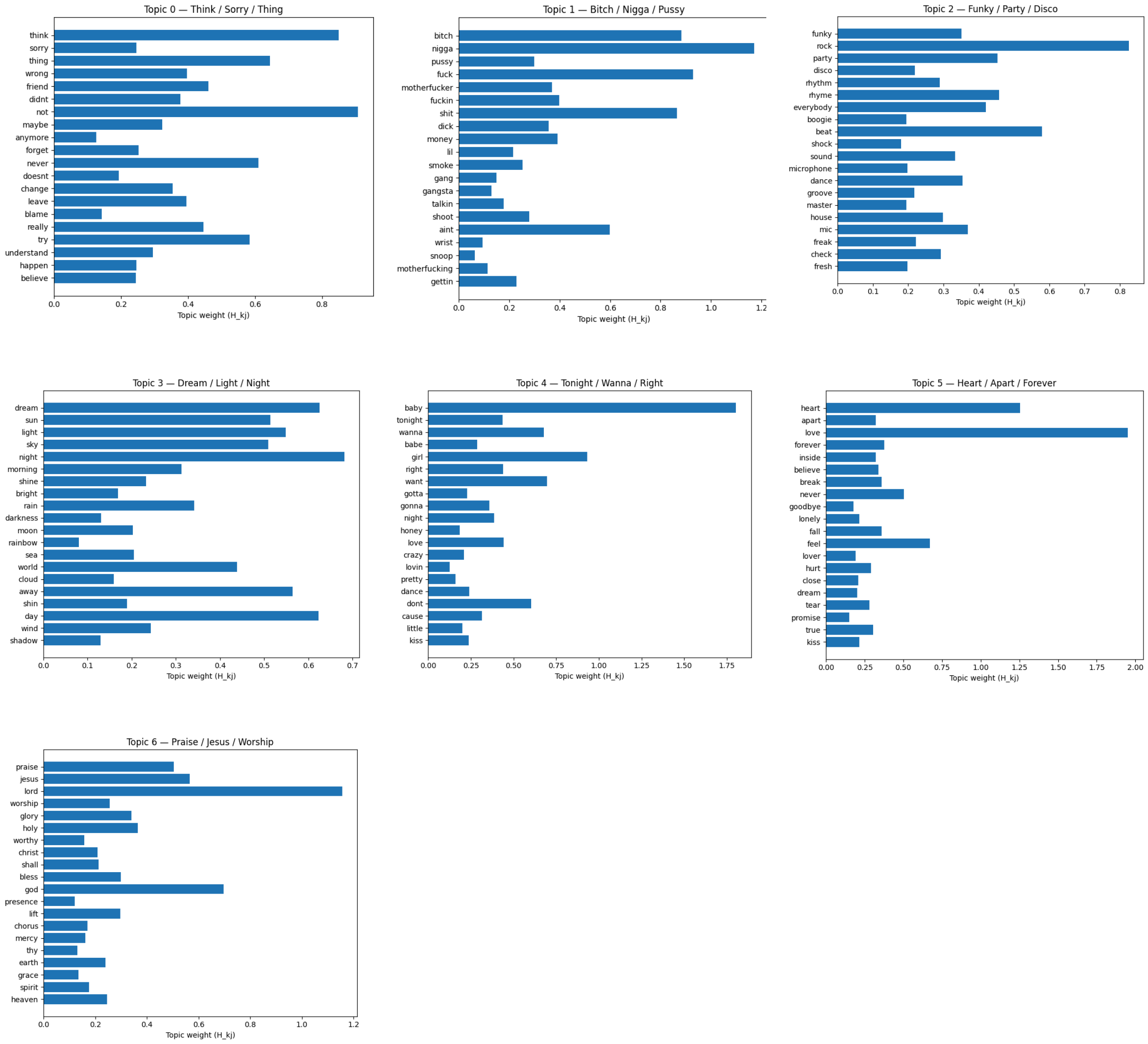
Appendix 7. Observed and fitted shares of positive sentiment lyrics for 2 genres



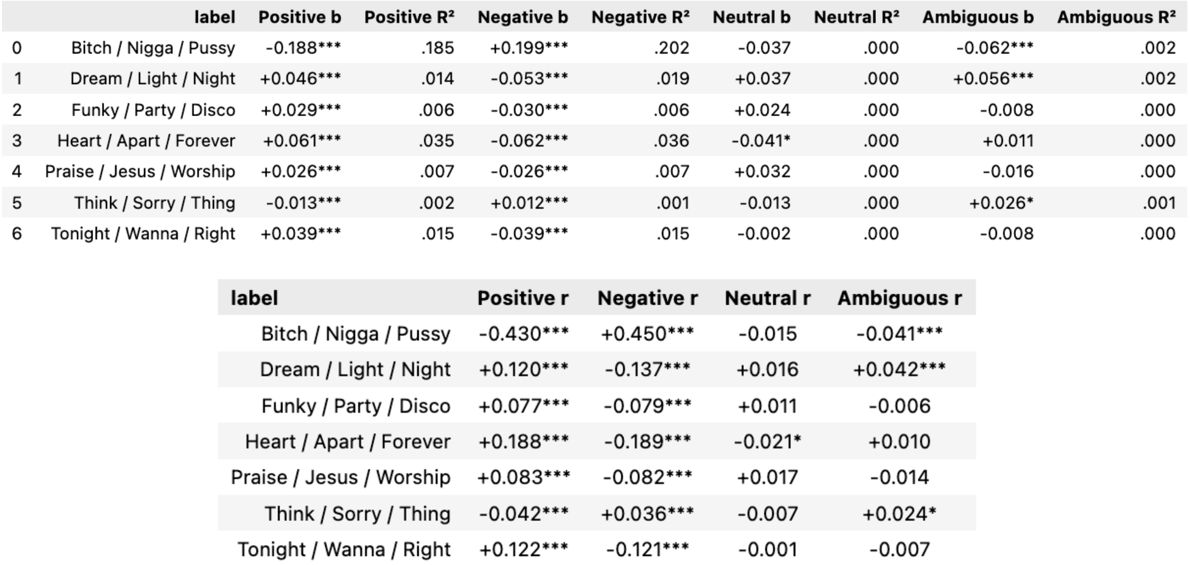
Appendix 8. Observed and fitted shares of negative sentiment lyrics for 2 genres



Appendix 9. Observed and fitted shares of polarity of sentiment for 2 genres



Appendix 10a-g. Distribution of terms with highest weights for 7 topics



	label	Positive b	Positive R^2	Negative b	Negative R^2	Neutral b	Neutral R^2	Ambiguous b	Ambiguous R^2
0	Bitch / Nigga / Pussy	-0.188***	.185	+0.199***	.202	-0.037	.000	-0.062***	.002
1	Dream / Light / Night	+0.046***	.014	-0.053***	.019	+0.037	.000	+0.056***	.002
2	Funky / Party / Disco	+0.029***	.006	-0.030***	.006	+0.024	.000	-0.008	.000
3	Heart / Apart / Forever	+0.061***	.035	-0.062***	.036	-0.041*	.000	+0.011	.000
4	Praise / Jesus / Worship	+0.026***	.007	-0.026***	.007	+0.032	.000	-0.016	.000
5	Think / Sorry / Thing	-0.013***	.002	+0.012***	.001	-0.013	.000	+0.026*	.001
6	Tonight / Wanna / Right	+0.039***	.015	-0.039***	.015	-0.002	.000	-0.008	.000

label	Positive r	Negative r	Neutral r	Ambiguous r
Bitch / Nigga / Pussy	-0.430***	+0.450***	-0.015	-0.041***
Dream / Light / Night	+0.120***	-0.137***	+0.016	+0.042***
Funky / Party / Disco	+0.077***	-0.079***	+0.011	-0.006
Heart / Apart / Forever	+0.188***	-0.189***	-0.021*	+0.010
Praise / Jesus / Worship	+0.083***	-0.082***	+0.017	-0.014
Think / Sorry / Thing	-0.042***	+0.036***	-0.007	+0.024*
Tonight / Wanna / Right	+0.122***	-0.121***	-0.001	-0.007

Appendix 11a-b. Statistical significance of topic-sentiment correlations for 7 topics