



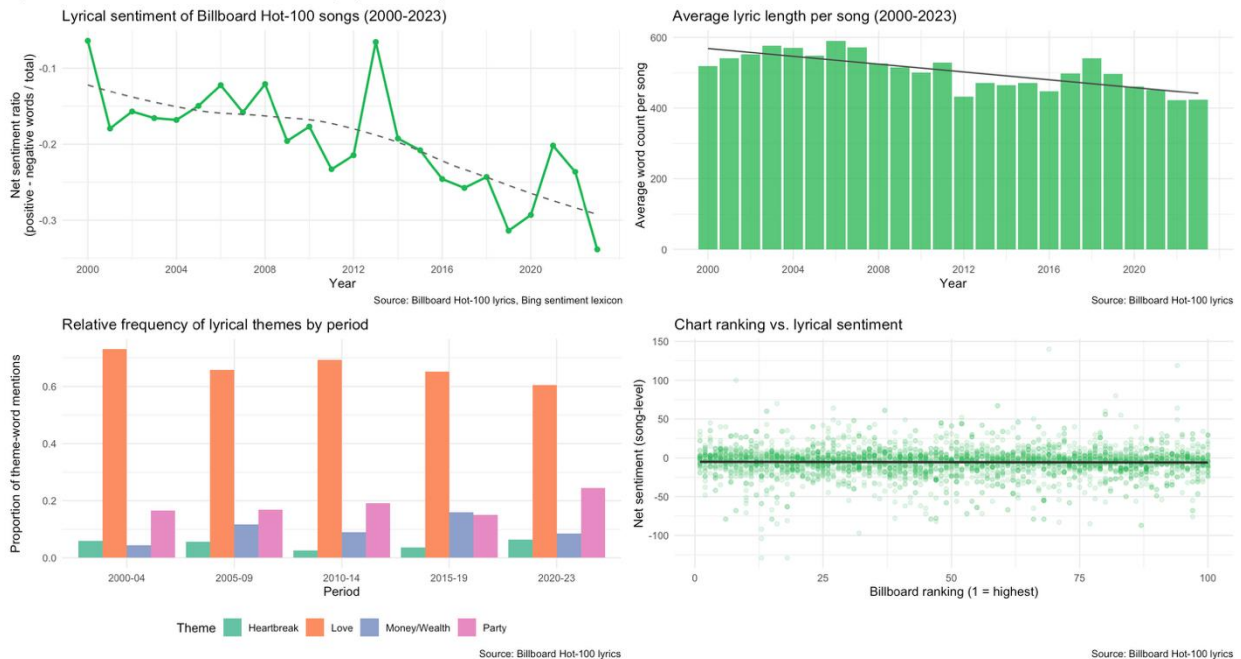
Lyrical Evolution and Chart Performance: Sentiment, Theme, and Complexity in Billboard Hot-100 Songs, 2000–2023

Module: IJC445

Word Count: 2,978

Registration Number: 250214208

Lyrical Evolution of Billboard Hot-100 Songs (2000-2023)



1. Knowledge Building

1.1 Topic and question

Popular music lyrics are increasingly treated as a large-scale record of cultural change, offering researchers a way to study shifting emotional norms at population scale (Brand et al., 2019). This report asks: how has the lyrical content of Billboard Hot-100 songs (sentiment, thematic focus, and lyrical density) evolved between 2000 and 2023, and does this evolution relate to chart performance? The question is narrowed to a single dataset, the Billboard Hot-100 from 2000 to 2023, so trends can be traced with the same consistency as, and compared against, prior large-scale studies of the same chart.

The question matters because several independent research groups, using different methods and time windows, have converged on the same pattern: popular music lyrics have become steadily more negative and structurally simpler over the past five to seven decades. Napier and Shamir (2018) found that anger, disgust, and sadness increased significantly in Billboard Hot-100 lyrics between 1951 and 2016, while joy and confidence declined. Brand et al. (2019) reached a similar conclusion using a larger, independent Billboard sample spanning 1965 to 2015, and found that chart position itself predicted lyrical negativity, with songs closer to the top of the charts containing more negative language, a pattern attributed to a content bias favouring negative material. Parada-Cabaleiro et al. (2024) and Varnum et al. (2021) documented a parallel decline in lyrical complexity, showing that vocabulary richness, readability, and structural variety have decreased while repetition has increased; Varnum et al. further showed simpler, more repetitive songs achieved higher chart positions. Most recently, Foramitti et al. (2025) confirmed both trends jointly using VADER sentiment and LZ77 compressibility on the

same dataset, while showing that societal shocks such as COVID-19 temporarily reversed rather than accelerated the negativity trend. Despite this convergence, no single visualisation has combined sentiment, theme, and lyrical density trends and tested their relationship to chart rank within one dataset, which is the gap this composite visualisation addresses.

1.2 Composite visualisation

Figure 1 (top-left) tracks net sentiment ratio of Hot-100 lyrics from 2000 to 2023 using the Bing sentiment lexicon. The trend is unmistakably downward, moving from a near-neutral ratio in 2000 to consistently negative values by the early 2020s, corroborating Napier and Shamir's (2018) and Foramitti et al.'s (2025) findings for the same chart over a longer window.

Figure 2 (top-right) shows average lyric word count per song by year. A mild but consistent decline is visible from the mid-2000s onward, aligning with Parada-Cabaleiro et al.'s (2024) and Varnum et al.'s (2021) documented rise in repetition and fall in lexical richness across popular music more broadly.

Figure 3 (bottom-left) breaks theme-word frequency (love, heartbreak, money, party) into five-year bins. Love dominates every period but declines proportionally over time, while party references rise sharply in 2020 to 2023, a thematic shift not previously reported for this dataset.

Figure 4 (bottom-right) plots chart ranking against song-level sentiment. Unlike Brand et al. (2019) and Varnum et al. (2021), who found real, if small, relationships between chart success and lyrical negativity or simplicity, this analysis shows an essentially flat relationship.

Taken together, the composite visualisation confirms that Billboard Hot-100 lyrics have grown more negative and structurally simpler between 2000 and 2023, consistent with decades of prior research on this chart. It also reveals that, for sentiment specifically, chart success does not track lyrical negativity as clearly as previous studies suggest, raising the question of whether more powerful statistical methods, rather than visual inspection alone, are needed to detect the subtler content-bias effects reported elsewhere in the literature, a limitation taken up further in Section 5.

2. Theoretical Frameworks

This section focuses on Figure 1, the sentiment-over-time line chart, and describes how it was developed using the ASSERT framework (Ferster, 2023; Ferster & Shneiderman, 2012) and explains its construction using Wilkinson's (2011) Grammar of Graphics (GoG).

Ask. The specific question behind Figure 1 was whether the emotional tone of Billboard Hot-100 lyrics has shifted over 2000 to 2023. Following Ferster's (2023) three-part query structure, this can be stated as: we are looking at Billboard Hot-100 lyrics because we want to find out how their sentiment has changed over time, in order for the audience to understand whether popular music has grown more negative. The intended audience is an academic reader already familiar with sentiment analysis, so the chart assumes some background knowledge and prioritises a clear overall trend over granular annotation, consistent with Ferster's (2023) guidance to define audience background and information needs before designing a visualisation.

Search. Lyrics and chart-rank data were sourced from a pre-compiled Billboard Hot-100 dataset combining Genius lyrics with chart metadata, a secondary rather than primary source in Ferster's (2023) terms, since neither the lyrics nor the rankings were collected first-hand for this project. Sentiment scoring relied on the Bing lexicon, a categorisation of English words into positive and negative classes; Napier and Shamir (2018) use a comparable dictionary-based approach on the same chart, which supports the reliability of this method as a known, precedented source of sentiment information for lyric analysis.

Structure. The underlying variables were year, a continuous interval-scale measure, and net sentiment ratio, a continuous ratio-scale measure bounded between negative one and one. These were derived by tokenising each song's lyrics into individual words, matching each word against the Bing lexicon, and aggregating positive and negative counts per year into a single ratio. This process converted unstructured raw text, which carries no inherent numerical structure, into a structured, one-row-per-year dataset suitable for plotting.

Envision. Several exploratory options were tested before settling on the final design, including plotting raw sentiment counts rather than a normalised ratio, and comparing a straight linear trend line against a loess-smoothed curve. The ratio and loess combination was retained because it revealed the overall trend pattern most clearly while dampening year-to-year noise, an example of the exploratory, iterative process Ferster (2023) describes as central to this stage of visualisation development.

Represent. In Wilkinson's (2011) terms, the chart's Variables are year and sentiment ratio; its Algebra involved aggregating individual word-level sentiment scores into a single yearly ratio, a summarising operation applied before plotting. Its Geometry combines a line geom, showing continuous change, with a point geom marking each year's exact value, while a second Statistics layer overlays a loess-smoothed regression curve to expose the underlying trend beneath annual fluctuation. The chart uses standard Cartesian Coordinates, with year on the x-axis and sentiment ratio on the y-axis.

Aesthetics were kept deliberately minimal: a single colour distinguishes the data series, and position along the y-axis, rather than colour or size, carries the main quantitative message, following Bertin's (1967) argument that position is the most legible visual variable for conveying precise quantitative differences.

Tell. The chart tells a single, linear story: lyrical sentiment has declined steadily over the study period. Framing the loess curve as a dashed, de-emphasised line alongside the more prominent solid trend line follows Ferster's (2023) view that effective visualisations should communicate one clear central message, with supporting detail visually subordinated rather than competing for the viewer's attention.

Together, these six stages show that Figure 1 was not simply plotted from available data, but shaped at each stage by a specific analytical question, a defined data structure, and an explicit set of Grammar of Graphics choices intended to make a single trend as legible as possible.

3. Accessibility

Accessibility in visualisation refers to the extent to which a chart's information can be perceived and understood by audiences with differing visual, cognitive, or sensory abilities, rather than assuming a single, typical viewer (Osiobe et al., 2024). This section focuses on Figure 3, the theme-frequency bar chart, and critiques its accessibility for two distinct audiences: people with colour vision deficiency (CVD), and screen reader users who cannot perceive the chart visually at all.

Figure 3 distinguishes its four theme categories, Heartbreak, Love, Money/Wealth, and Party, using RColorBrewer's Set2 palette, a common default choice in ggplot2. To test whether this choice is genuinely accessible rather than assumed to be, the palette's four colours were passed through a colour vision deficiency simulation covering deuteranomaly, protanomaly, and tritanomaly, the three principal forms of CVD (Culp, 2012). The results reveal a real design flaw: under simulated protanomaly, Money/Wealth and Party become almost indistinguishable, differing by a colour distance of only 7.7 in simulated RGB space, well below the threshold generally considered safely distinguishable. Under deuteranomaly, the most common form of CVD, both Heartbreak versus Party and Money/Wealth versus Party fall into a similarly confusable range. This means that, for a meaningful proportion of readers with red-green colour deficiency, the chart's four supposedly distinct categories effectively collapse into three or fewer, undermining its core communicative purpose.

This finding echoes Culp's (2012) distinction between inherited CVD, most commonly red-green confusion, and acquired CVD, more often blue-green confusion associated with ageing and conditions such as diabetes; a palette tested against only one type, as many default software palettes are, risks excluding readers with the other. Following Osiobe et al.'s (2024) recommendations, a straightforward improvement would be to supplement colour with a second, redundant visual cue, such as distinct bar patterns or hatching for each theme, so that category identity does not depend on colour discrimination alone. Alternatively, tools such as the cblindplot package (Rocchini et al., 2023), designed specifically for colourblind-safe plotting in R, could have been used to generate a palette pre-validated against multiple CVD types rather than relying on a general-purpose default.

Figure 3 is also inaccessible to screen reader users in a more fundamental sense: as a static PNG image embedded in a Word document, it offers no alternative textual or structural representation at all. Sharif et al. (2021) found that screen reader users extract information 61.48% less accurately and spend 210.96% more time interacting with online visualisations compared to sighted users, and that many charts are simply undetectable to screen readers in the first place. Zong et al. (2022) further argue that meaningful accessibility requires attention to a chart's structure, navigation, and description, not merely the presence of alt text; by these standards, Figure 3 currently offers only the lowest tier of accessibility, since no alt text, data table, or structural markup accompanies it. A minimally accessible improvement would be to pair the chart with a simple summary data table listing theme proportions by period, giving screen reader users direct access to the underlying values even without a navigable graphical structure.

Taken together, Figure 3 illustrates that a chart can appear well-designed by conventional standards while still failing two substantial audiences: colour is used as the sole encoding for a categorical variable without CVD testing, and no non-visual alternative exists for readers who cannot see the chart at all.

4. Visualisation Choice

4.1 Justification of the type of visualisation

This section focuses on Figure 4, the scatter plot of chart ranking against song-level sentiment. The goal of Figure 4 is to reveal whether more successful (higher-ranked) songs contain systematically different lyrical sentiment than less successful ones, a genuinely open empirical question given that Brand et al. (2019) and Varnum et al. (2021) report small but real relationships between chart success and lyrical content using more powerful statistical models than a single visualisation can apply. A scatter plot was chosen because it is the standard visual form for examining the relationship between two continuous variables, here ranking and net sentiment, without imposing any prior aggregation or binning on the data.

This choice is well supported by Cleveland and McGill's (1984, 1985) theory of graphical perception, which ranks position along a common scale as the single most accurately decoded visual task. A scatter plot encodes both variables through position alone, ranking on the horizontal axis and sentiment on the vertical, meaning every individual data point can in principle be read with the highest available perceptual accuracy, rather than through less reliable channels such as colour or area. Overlaying a linear trend line further exploits this accuracy: rather than asking the viewer to mentally average roughly 3,400 overlapping points, the fitted line performs that aggregation explicitly, leaving only the comparatively easier task of judging the line's slope. In Shneiderman's (2003) task-based taxonomy, this corresponds most closely to the relate task, in which a visualisation's central purpose is to expose or test an association between two variables rather than to support overview, filtering, or detail lookup individually.

4.2 Alternatives

Two alternative chart types were considered and rejected. The first is a two-dimensional density plot, such as a hexbin chart, in which the plotting area is divided into a grid of hexagonal cells shaded by the number of points they contain. This would directly solve the scatter plot's most visible weakness: with almost 3,400 individual songs, points cluster so heavily near zero sentiment that many overlap entirely, even with transparency applied. A hexbin plot would make the density of that cluster immediately visible rather than implied. However, it encodes the count information through colour saturation, one of the least accurately perceived channels in Cleveland and McGill's (1984) ranking, meaning that while overall density becomes clearer, the precise sentiment value of any individual song becomes unrecoverable, a significant loss given that the dataset's outliers (songs with unusually high or low sentiment) are arguably the most analytically interesting points.

The second alternative is a box plot of sentiment grouped into ranking bands, for example top-10, 11 to 50, and 51 to 100. This would resolve overplotting entirely and would let a reader compare medians and interquartile ranges directly using position along a common scale, again a highly accurate perceptual task. Its weakness is that it imposes arbitrary

boundaries on a continuous ranking variable and discards all point-level detail, making it impossible to detect a genuine but weak linear trend of the kind Varnum et al. (2021) report, since any real but small effect would be flattened by aggregating hundreds of songs into only two or three summary boxes.

The scatter plot with an overlaid trend line was retained over both alternatives because the analytical goal of Figure 4 is specifically to test for a subtle, potentially small relationship, the kind that a hexbin plot's colour encoding would obscure and a box plot's binning would risk erasing entirely, even at some cost to visual clarity from overplotting.

5. Implications and Improvements

5.1 Ethical implication

This section focuses on Figure 2, the average lyric word count per song by year. Unlike some of the classic distortion techniques catalogued by Pandey et al. (2015), such as truncated or inverted axes, Figure 2 avoids these specific pitfalls: its y-axis begins at zero and no scale is manipulated. By the narrow definition of visual deception used in early misinformation research, the chart is technically faithful to the underlying data.

However, Lisnic et al. (2023) show that most real-world visualisation misinformation does not come from such technical distortions at all: in their analysis of thousands of COVID-19 chart posts, only around 12% involved design violations, while the great majority of misleading visualisations were faithfully plotted but vulnerable to being misread through omitted context, cherry-picked time frames, or unwarranted causal inference. Figure 2 fits this description of a vulnerable rather than a deceptive visualisation. Presented alone, its declining trend line invites an implicit normative reading, that lyrics are becoming worse, simpler, or less sophisticated, when the chart itself only documents a change in average word count. This is precisely the kind of unstated premise Lisnic et al. (2023) identify as central to weak inductive reasoning from charts: the visualisation does not claim this evaluative conclusion, but a viewer supplies it themselves if the chart is not framed carefully. Cairo (2014) argues that evidence-driven visual communicators bear a responsibility not merely to plot data accurately but to anticipate how an unqualified chart might be read; a purely descriptive downward trend, shown without comment, risks exactly this kind of unintended interpretive leap.

A further, more concrete risk concerns time-frame omission, and checking for it directly confirms the concern rather than leaving it hypothetical. Segmenting this dataset's word-count trend at 2016 shows a genuine local reversal: average word count rose sharply from 2016 to 2018 (approximately 47 words per year) before resuming an even steeper decline from 2018 to 2023 (approximately 23 words per year), while the overall pre- and post-2016 linear slopes look similar once averaged across the full period. A single smoothed trend line spanning 2000 to 2023, of the kind used in Figure 2, therefore does mask this local uptick, loosely corroborating Foramitti et al.'s (2025) report of a complexity increase coinciding with the years 2017 to 2021, though in this dataset the reversal is shorter-lived. This is precisely the kind of omission Lisnic et al. (2023) describe as a time-frame cherry-picking vulnerability, arising here not from any deliberate exclusion of data but from the smoothing technique itself.

5.2 Future improvements

Three improvements would meaningfully strengthen Figure 2. First, word count is a crude proxy for lyrical complexity: it captures length but not repetition, vocabulary richness, or readability. Future iterations should replace it with an established compression-based measure such as the LZ77 compressibility ratio used by Varnum et al. (2021) and

Foramitti et al. (2025), which more directly captures the repetitiveness that these studies argue is the real driver of perceived simplicity, rather than length alone.

Second, the chart should be explicitly annotated to prevent the unwarranted normative reading discussed above. Adding a brief caption noting that word count is a purely descriptive, non-evaluative measure, and marking any known inflection points such as the post-2016 reversal reported elsewhere in the literature, would close the gap between what the chart shows and what a viewer might otherwise infer.

Third, the accessibility limitations identified for Figure 3 in Section 3 apply equally here: Figure 2 currently exists only as a static image with no accompanying data table or alt text, making it unreadable to screen reader users. Pairing it with a simple yearly summary table, alongside a colourblind-safe palette check of the kind performed for Figure 3, would extend accessibility improvements consistently across the whole composite visualisation rather than treating each chart in isolation.

References

1. Bertin, J. (1967). *Sémiologie graphique: Les diagrammes les réseaux, les cartes*. Paris: Gauthier-Villars.
2. Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable?. *IEEE transactions on visualization and computer graphics*, 19(12), 2306-2315.
3. Brand, C. O., Acerbi, A., & Mesoudi, A. (2019). Cultural evolution of emotional expression in 50 years of song lyrics. *Evolutionary Human Sciences*, 1, e11.
4. Cairo, A. (2014). Graphics lies, misleading visuals: Reflections on the challenges and pitfalls of evidence-driven visual communication. In *New challenges for data design* (pp. 103-116). London: Springer London.
5. Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387), 531-554.
6. Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828-833.
7. Culp, G. M. (2012). Increasing accessibility for map readers with acquired and inherited colour vision deficiencies: a re-colouring algorithm for maps. *The cartographic journal*, 49(4), 302-311.
8. Ferster, B. (2023). *Interactive visualization: Insight through inquiry*. MIT Press.
9. Foramitti, M., Nater, U. M., Lamm, C., & Martins, M. (2025). Societal crises disrupt long-term increases in stress, negativity, and simplicity in US Billboard song lyrics from 1973 to 2023. *Scientific Reports*, 15(1), 41733.
10. Lisnic, M., Polychronis, C., Lex, A., & Kogan, M. (2023, April). Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-21). ACM.
11. Napier, K., & Shamir, L. (2018). Quantitative sentiment analysis of lyrics in popular music. *Journal of Popular Music Studies*, 30(4), 161-176.
12. Osiobe, E. U. (2024). Enhancing data visualization accessibility: a case for equity and inclusion. *Engineering and Technology Quarterly Reviews*, 7(1), 24–32.
13. Pandey, A. V., Rall, K., Satterthwaite, M. L., Nov, O., & Bertini, E. (2015, April). How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 1469-1478). ACM.
14. Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., & Zangerle, E. (2024). Song lyrics have become simpler and more repetitive over the last five decades. *Scientific Reports*, 14(1), 5531.
15. Rocchini, D., Nowosad, J., D’Introno, R., Chieffallo, L., Bacaro, G., Gatti, R. C., ... & Thouverai, E. (2023). Scientific maps should reach everyone: The cblindplot R

package to let colour blind people visualise spatial patterns. *Ecological Informatics*, 76, 102045.

16. Sharif, A., Chintalapati, S. S., Wobbrock, J. O., & Reinecke, K. (2021, October). Understanding screen-reader users' experiences with online data visualizations. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 1-16). ACM.
17. Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization* (pp. 364-371). Morgan Kaufmann.
18. Varnum, M. E., Krems, J. A., Morris, C., Wormley, A., & Grossmann, I. (2021). Why are song lyrics becoming simpler? A time series analysis of lyrical complexity in six decades of American popular music. *PloS one*, 16(1), e0244576.
19. Wilkinson, L. (2011). The grammar of graphics. In *Handbook of computational statistics: Concepts and methods* (pp. 375-414). Berlin, Heidelberg: Springer Berlin Heidelberg.
20. Zong, J., Lee, C., Lundgard, A., Jang, J., Hajas, D., & Satyanarayan, A. (2022, June). Rich screen reader experiences for accessible data visualization. In *Computer Graphics Forum* (Vol. 41, No. 3, pp. 15-27).

Appendix: R codes

```
# =====  
# IJC445 Data Visualisation - Composite Visualisation  
# Topic: Evolution of lyrical content in Billboard Hot-100 songs (2000-2023)  
# and its relationship with chart performance  
# Dataset: Billboard Hot-100 (2000-2023) with Genius lyrics  
# =====  
  
# --- 1. Load packages -----  
library(readxl) # read the .xlsx data  
library(dplyr) # data wrangling  
library(tidyr) # reshaping  
library(stringr) # text cleaning  
library(tidytext) # tokenisation and sentiment lexicons  
library(ggplot2) # visualisation  
library(patchwork) # combine charts into a composite panel  
  
# --- 2. Load data -----  
raw_data <- read_excel("Spotify_data.xlsx", sheet = "billboard_24years_lyrics_spotify")  
  
# --- 3. Clean lyrics -----  
# The scraped lyrics contain two Genius artefacts that must be removed before  
# any text analysis, otherwise word counts and sentiment scores are distorted:  
# (a) "You might also like" - a Genius ad insertion appearing mid-lyric  
# (b) a trailing "<number>Embed" tag appended to the very end of the string  
clean_lyrics <- raw_data %>%  
  mutate(  
    lyrics_clean = lyrics %>%  
      str_remove_all("You might also like") %>%  
      str_remove("\\d+Embed\\s*$") %>%  
      str_squish()  
  ) %>%  
  filter(!is.na(lyrics_clean), lyrics_clean != "")  
  
# --- 4. Tokenise lyrics into words (one row per word per song) -----  
lyrics_words <- clean_lyrics %>%  
  select(song, band_singer, year, ranking, lyrics_clean) %>%  
  mutate(song_id = row_number()) %>%  
  unnest_tokens(word, lyrics_clean) %>%  
  anti_join(stop_words, by = "word") # remove common stop words (the, and, ...)  
  
# Word count per song (before stop-word removal, to reflect true lyric length)  
word_counts_per_song <- clean_lyrics %>%  
  mutate(song_id = row_number()) %>%  
  select(song_id, year, ranking, lyrics_clean) %>%  
  mutate(n_words = str_count(lyrics_clean, "\\S+"))
```

```
# ---- 5. Chart 1: Average sentiment per year (Bing lexicon) -----
```

```
bing <- get_sentiments("bing") %>% distinct(word, .keep_all = TRUE)
```

```
sentiment_by_year <- lyrics_words %>%
```

```
  inner_join(bing, by = "word") %>%
```

```
  count(year, sentiment) %>%
```

```
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
```

```
  mutate(net_sentiment = positive - negative,
```

```
         total = positive + negative,
```

```
         sentiment_ratio = net_sentiment / total) # normalised for comparability across years
```

```
fig1 <- ggplot(sentiment_by_year, aes(x = year, y = sentiment_ratio)) +
```

```
  geom_line(colour = "#1DB954", linewidth = 1) +
```

```
  geom_point(colour = "#1DB954", size = 1.8) +
```

```
  geom_smooth(method = "loess", se = FALSE, colour = "grey40", linetype = "dashed", linewidth = 0.6) +
```

```
  scale_x_continuous(breaks = seq(2000, 2023, 4)) +
```

```
  labs(
```

```
    title = "Lyrical sentiment of Billboard Hot-100 songs (2000-2023)",
```

```
    x = "Year", y = "Net sentiment ratio\n(positive - negative words / total)",
```

```
    caption = "Source: Billboard Hot-100 lyrics, Bing sentiment lexicon"
```

```
  ) +
```

```
  theme_minimal(base_size = 11)
```

```
# ---- 6. Chart 2: Average word count per year -----
```

```
wordcount_by_year <- word_counts_per_song %>%
```

```
  group_by(year) %>%
```

```
  summarise(avg_words = mean(n_words, na.rm = TRUE), .groups = "drop")
```

```
fig2 <- ggplot(wordcount_by_year, aes(x = year, y = avg_words)) +
```

```
  geom_col(fill = "#1DB954", alpha = 0.8) +
```

```
  geom_smooth(method = "lm", se = FALSE, colour = "grey30", linewidth = 0.6) +
```

```
  scale_x_continuous(breaks = seq(2000, 2023, 4)) +
```

```
  labs(
```

```
    title = "Average lyric length per song (2000-2023)",
```

```
    x = "Year", y = "Average word count per song",
```

```
    caption = "Source: Billboard Hot-100 lyrics"
```

```
  ) +
```

```
  theme_minimal(base_size = 11)
```

```
# ---- 7. Chart 3: Theme word frequency across time periods -----
```

```
# Theme word dictionaries used to classify lyrics into four categories
```

```
theme_words <- tribble(
```

```
  ~word,    ~theme,
```

```
  "love",   "Love",
```

```
  "heart",  "Love",
```

```
  "baby",   "Love",
```

```
  "money",  "Money/Wealth",
```

```

"cash", "Money/Wealth",
"rich", "Money/Wealth",
"party", "Party",
"dance", "Party",
"night", "Party",
"cry", "Heartbreak",
"tears", "Heartbreak",
"lonely", "Heartbreak"
)

lyrics_words_period <- lyrics_words %>%
  mutate(period = cut(year, breaks = c(1999, 2004, 2009, 2014, 2019, 2024),
    labels = c("2000-04", "2005-09", "2010-14", "2015-19", "2020-23"),
    include.lowest = TRUE))

theme_freq <- lyrics_words_period %>%
  inner_join(theme_words, by = "word") %>%
  count(period, theme) %>%
  group_by(period) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()

fig3 <- ggplot(theme_freq, aes(x = period, y = prop, fill = theme)) +
  geom_col(position = "dodge") +
  labs(
    title = "Relative frequency of lyrical themes by period",
    x = "Period", y = "Proportion of theme-word mentions", fill = "Theme",
    caption = "Source: Billboard Hot-100 lyrics"
  ) +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal(base_size = 11) +
  theme(legend.position = "bottom")

# ---- 8. Chart 4: Ranking vs sentiment -----
song_level_sentiment <- clean_lyrics %>%
  mutate(song_id = row_number()) %>%
  select(song_id, ranking, year, lyrics_clean) %>%
  unnest_tokens(word, lyrics_clean) %>%
  anti_join(stop_words, by = "word") %>%
  inner_join(bing, by = "word") %>%
  count(song_id, ranking, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(net_sentiment = positive - negative)

fig4 <- ggplot(song_level_sentiment, aes(x = ranking, y = net_sentiment)) +
  geom_point(alpha = 0.15, colour = "#1DB954") +
  geom_smooth(method = "lm", colour = "grey20", se = TRUE) +
  labs(

```

```

    title = "Chart ranking vs. lyrical sentiment",
    x = "Billboard ranking (1 = highest)", y = "Net sentiment (song-level)",
    caption = "Source: Billboard Hot-100 lyrics"
  ) +
  theme_minimal(base_size = 11)

# ---- 9. Composite visualisation -----
composite <- (fig1 | fig2) / (fig3 | fig4) +
  plot_annotation(
    title = "Lyrical Evolution of Billboard Hot-100 Songs (2000-2023)",
    theme = theme(plot.title = element_text(size = 14, face = "bold"))
  )

# Save the composite visualisation as a PNG file
ggsave("composite_visualisation.png", composite, width = 12, height = 8, dpi = 300)

print(composite)

```