Author
Tobias Wollendorfer
K12008678

Submission
Institute for
Computational Perception

Thesis Supervisor
Prof. Dr. **Markus Schedl**

September 2025

Sentiment and Linguistic Patterns in YouTube Music Video Comments Across Genres

Bachelor's Thesis

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Contents

Abstract

Music is closely linked to emotional experience and the emergence of social communities. In the digital age, Platforms such as YouTube have become central spaces for music consumption and discussion. YouTube comments in particular provide a large-scale record of spontaneous audience reactions, offering valuable insights into how listeners across different musical genres express themselves online.

This thesis investigates sentiment and linguistic patterns in YouTube comments on music videos, with a focus on differences between genres. A representative dataset was constructed by combining the JKU dataset of YouTube music links with the Music4All-Onion dataset for genre classification. Using a conflict-resolution algorithm, 50 representative videos were selected for each of 260 genres, yielding over 11,000 videos. From these, 100 recent comments per video were collected via the YouTube Data API, resulting in a dataset of 356,973 English comments across 233 genres. Each comment was preprocessed and analyzed using VADER for sentiment scoring and LIWC for psycholinguistic feature extraction.

The results highlight clear genre-specific differences in sentiment. Spiritual and smooth genres consistently received the most positive comments, while hip-hop and aggressive genres showed the lowest sentiment scores despite high engagement. Polarization analyses revealed that aggressive and rap-related styles exhibit the strongest coexistence of enthusiasm and critique, while worship and jazz genres are more uniformly positive. Finally, correlations between LIWC dimensions and VADER sentiment confirmed the validity of both tools, with expected associations (e.g., positive emotion ↔ VADER positive, swear words ↔ VADER negative).

Chapter 1

Introduction

1.1 Motivation

Music is one of the most powerful and universal forms of human expression. It shapes cultural identity, regulates emotions, and forms social connection across communities. In the digital era, music consumption has shifted to online platforms, with YouTube emerging as the largest global centre for music streaming and community discussions. With more than 1.2 billion monthly users, YouTube not only gives people access to music but also provides them with the chance to present feedback in the form of comments, likes, and shares.

These comments represent a unique opportunity to study audience engagement and genre-specific communication styles. Unlike professional reviews, they capture spontaneous, authentic reactions from listeners across diverse genres and backgrounds. Understanding how these reactions differ across genres provides insight into how music communities construct meaning, express emotions, and build social connections.

However, analyzing such data is challenging. Social media comments are highly informal, often including slang or irony, which complicates interpretation. Traditional sentiment analysis methods—such as lexicon-based models—may fail to capture the nuances of genre-specific language, particularly in culturally distinct communities such as hip-hop. This thesis addresses these challenges by combining computational sentiment analysis with psycholinguistic feature extraction, aiming to reveal how music genres differ in their online communication.

1.2 Objectives and Research Questions

The main objective of this thesis is to investigate how sentiment and linguistic patterns vary across YouTube music video comments, and how these patterns relate to genre characteristics and audience engagement. Based on this objective, the following research questions are addressed:

1. **Sentiment and Engagement/Popularity:** How does the sentiment of YouTube music video comments relate to audience engagement and popularity, as reflected in metrics such as likes, views, and comment counts?
2. **Sentiment Differences Across Genres:** How do average sentiment scores differ across musical genres, and which genres exhibit the most positive, negative, or neutral patterns?
3. **Polarization:** To what extent are comments polarized, both within individual messages and across entire genres, and what does this reveal about community communication styles?

4. **Psycholinguistic Features:** Which psycholinguistic features (e.g., swearing, social words, conflict terms) distinguish genres, and how do these features interact with sentiment expression?

These research questions guide the dataset construction, analysis design, and interpretation of results presented in the following chapters.

Chapter 2

Related Work

Social Media Comments Analysis

The analysis of social media comments has become an increasingly important research field. Most existing studies have focused on platforms such as Twitter and Meta-platforms, including Facebook and Instagram. With more than 1.2 billion monthly users, YouTube is currently the most visited social media platform in the world and the second most visited website overall after Google Search [13].

Research on YouTube comment analysis has primarily focused on sentiment analysis, spam detection, and user engagement patterns. Recent studies have proposed advanced methods for filtering spam content from comments [11].

Another growing research field is toxicity and hate-speech detection. Hartvigsen et al. introduced TOXIGEN, a large-scale dataset of machine-generated and human-written statements designed to capture both explicit and implicit toxic language. Their results showed that models fine-tuned on TOXIGEN outperform baselines in detecting subtle forms of hate speech [5].

VADER-based Sentiment Analysis

Many researchers use VADER for analyzing short, informal, and emoji-rich social media text because it handles intensifiers, negation, punctuation and emoticons directly in its rule set [8]. Beyond the original Twitter-style validation, VADER has been widely applied to YouTube comments to quantify audience polarity. For instance, Chalkias *et al.* (2023) examined 167,987 comments from educational YouTube channels using both VADER and TextBlob, finding that neutral sentiment dominates and that VADER tends to report more neutral statements compared to TextBlob [4]. Another recent study by Zhang (2025) compared VADER and TextBlob on approximately 18,000 YouTube video comments and correlated sentiment with likes and views, showing that while both tools provide useful signals, TextBlob produced more stable correlations in this context [14]. In these settings, the compound score serves as a practical and interpretable score for overall valence, enabling large-scale comparisons between topics and channels.

LIWC-based Linguistic Profiling

In addition to valence detection, LIWC provides psychologically grounded categories spanning affect, social processes, cognitive mechanisms, authenticity, and topical concerns. The approach has been extensively validated and used to profile communication styles and psychological correlates in natural language across online platforms [10, 12].

In the context of YouTube, LIWC has been used to uncover community norms, such as social bonding, swearing/hostility markers, or religiosity, that are not captured by sentiment alone, thus offering a richer view of how users express themselves. For example, Chae *et al.* (2024) analyzed COVID-19 video transcripts and comments from medical YouTubers using LIWC categories including analytical thinking and emotion (anxiety, anger, sadness), finding that these linguistic/emotional dimensions are associated with viewer engagement and emotional alignment between channel creators and audiences [3].

Why Combine VADER and LIWC?

VADER and LIWC answer complementary questions. VADER provides a robust social media-tuned estimate of the valence of comments (positive / negative / neutral) with a single comparable compound score suitable for large-scale analyses [8]. LIWC, decomposes language into psychologically meaningful dimensions (e.g., *Affect*, *Swear*, *Social*, *Cognition*), allowing one to interpret *how* communities communicate, not just *how positive or negative* they are [10].Combining both methods makes it possible to map sentiment hierarchies across genres, explain those differences through concrete linguistic markers, and relate community endorsement (likes) to either overall valence or specific stylistic and psycholinguistic features.

Music-related Social Media Research

Music-related videos on YouTube generate distinct engagement behaviors compared to other types of content. Research in music psychology has shown that music listening is driven by specific psychological functions such as mood regulation, identity expression, and social connection. These motivations strongly shape how listeners engage with and respond to music content. Listeners actively select music according to explicit listening intentions and seek playlists aligned with their goals, as shown by the ExIM study [6]. Similarly, Arif et al. (2024) conducted a content analysis of BTS music video comments and found strong evidence of parasocial interactions, with fans frequently expressing authenticity, affection, and social bonding with the artist, highlighting the social dimension of the music comment culture [1].

Extending these insights to YouTube more broadly, comments on music videos often contain explicit emotional expressions and personal accounts of why users listen to specific songs. This suggests that the psychological drivers of music listening identified in controlled studies can also be observed directly in user-generated content.

Bauer and Schedl (2019) further contributed to understanding genre-based behavior by introducing the concept of global and country-specific *mainstreaminess.* Using large-scale Last.fm data, they defined measures to capture how closely individual user preferences align with global or local popularity trends. Their results revealed strong cross-country variation, with some regions showing alignment with international mainstream genres, while others emphasized localized and niche listening patterns. This demonstrates that genre preferences are not only individual but also shaped by cultural and community-level contexts [2].

Summary

These studies show that music-related comments on YouTube provide more than casual reactions. They reflect emotional transmission, parasocial interaction, and culturally shaped genre preferences. Building on these insights and leveraging the complementary strengths of VADER and LIWC, this thesis analyzes YouTube music video comments to examine how sentiment and linguistic patterns vary across musical genres and how different communities express themselves within these online spaces.

Chapter 3

Dataset

3.1 Data Sources

This study combines three primary data sources to create a comprehensive dataset for analyzing sentiment patterns across music genres. The **JKU Dataset** (*id_ youtube_ url.csv*) provides unique identifiers paired with YouTube URLs linking to music videos. The Music4All-Onion dataset [9] connects these songs to genres based on audio features, enabling systematic genre classification. Finally, the **YouTube Data API v3** serves as the primary collection mechanism for user-generated comments and engagement metadata.

These sources were selected for their complementary strengths: the JKU Dataset ensures broad musical coverage, the Music4All-Onion Dataset provides objective, audio-feature-based genre classifications avoiding subjective labeling bias, and YouTube represents the world's largest music streaming platform with authentic user interactions.

3.2 Representative Song Selection

The primary objective was to achieve fair representation across all music genres while avoiding bias toward larger, more popular genres that typically dominate music datasets. Larger genres often exhibit more cross-genre overlap and achieve higher relevance scores, which can marginalize underrepresented musical genres.

The selection algorithm was therefore designed around three key principles: (i) ensuring exactly k unique videos per genre (preventing assignment of the same video to multiple genres), (ii) implementing fair conflict resolution when videos were contested by multiple genres, and (iii) maximizing representation of smaller genres through a fallback scoring mechanism that considers alternative options when conflicts arise.

In practice, the algorithm processed 260 initial genres, assigned the top-scoring videos to each genre, detected and resolved conflicts using fallback scores (where genres with weaker alternatives retained contested videos), and refilled losing genres with their next-best unassigned candidates. This iterative process continued until convergence was reached.

The outcome yielded 50 representative videos per genre across all qualifying genres, ensuring balanced representation while preserving video uniqueness in the final dataset. A detailed step-by-step description of the algorithm is provided in Appendix A.

3.3 Comment Collection & Metadata

Using the YouTube Data API v3, up to 100 comments were collected per representative video, targeting the most recent user interactions. This systematic approach ensured uniform sampling across all genres regardless of their individual popularity or engagement levels.

The collection process captured metadata at both the comment and video levels. Comment-level information included unique identifiers, authorship details, timestamps, engagement indicators such as likes, and reply structures. Video-level metadata comprised descriptive attributes (title, description, publication date), content characteristics (duration, definition, captions), engagement statistics (views, likes, comments, favorites), and technical properties (licensing status, privacy settings, and platform-specific flags).

This two-level metadata structure allows for the joint analysis of individual comment characteristics and the broader video context, providing a comprehensive basis for examining how genre-specific factors influence audience engagement. In total, the dataset contains 161 distinct features for each comment.

3.4 Preprocessing Pipeline

3.4.1 Language Filtering

Given that both VADER sentiment analysis and LIWC require English text for accurate processing, only English comments were retained for analysis. This decision, while necessary, introduces a cultural bias by excluding non-English speaking music communities and potentially underrepresenting genres popular in non-English speaking regions. The language filtering reduced the dataset from 620,736 initial comments to 356,973 English comments.

3.4.2 Text Standardization

A systematic text preprocessing pipeline standardized comment text through three stages: whitespace normalization (removing excess spaces, tabs, and line breaks) and case normalization (converting to lowercase for consistent analysis). This standardization ensures consistent input for the sentiment and linguistic analysis tools.

3.4.3 Genre Filtering

To ensure statistical reliability, genres with fewer than 500 English comments were excluded from analysis. This threshold balances the need for meaningful sample sizes with overall dataset coverage, reducing the genre count from 260 to 233 while keeping the analysis meaningful. The 27 excluded genres represent primarily niche or non-English-dominant musical styles.

3.5 Feature Extraction

3.5.1 Sentiment Analysis

VADER (Valence Aware Dictionary for sentiment Reasoning) was selected for sentiment analysis due to its specific optimization for social media text and its ability to handle informal language elements such as slang and emojis commonly found in online communication [8]. VADER generates four sentiment scores per comment: positive, negative, and neutral proportions, plus a compound score normalized between -1 (most negative) and +1 (most positive) that serves as the primary sentiment indicator [7].

3.5.2 Linguistic Features

LIWC (Linguistic Inquiry and Word Count) was employed to extract psycholinguistic features across multiple dimensions: psychological processes (cognitive, emotional, social), personal concerns (work, money, religion), linguistic dimensions (word count, pronouns, articles), and grammatical categories (verbs, adjectives, prepositions). These features enable analysis of communication styles and psychological patterns beyond basic sentiment.

3.6 Final Dataset Characteristics

The final dataset comprises 356,973 English comments distributed across 233 music genres, with each genre containing a minimum of 500 comments to ensure statistical validity. Despite efforts to minimize temporal dispersion by scraping only the most recent comments, the dataset still covers multiple years of YouTube activity, providing valuable insight into temporal dynamics in music communication.

Genre representation varies significantly, with popular genres like pop and rock contributing thousands of comments while niche genres approach the 500-comment minimum threshold.

The preprocessing pipeline successfully standardized textual content while preserving essential metadata, creating a robust foundation for comparative sentiment and linguistic analysis across diverse musical communities.

Chapter 4

Methods

This chapter explains the analysis design and statistical methods. The actual results and figures are presented in Chapter 5.

4.1 Sentiment Scoring (VADER)

We quantify comment-level sentiment using VADER (Valence Aware Dictionary and sEntiment Reasoner) [8], a lexicon- and rule-based tool designed for short, informal, and emoji-rich social media text. Each comment receives four sentiment indicators:

- `positive` $\in [0, 1]$: proportion of text conveying positive valence,
- `negative` $\in [0, 1]$: proportion conveying negative valence,
- `neutral` $\in [0, 1]$: proportion conveying neutral valence,
- `compound` $\in [-1, 1]$: normalized overall sentiment score.

The proportional scores sum to one, and the compound score summarizes the text's overall polarity after applying VADER's heuristics (e.g., negation, degree modifiers, punctuation, capitalization).

4.2 Psycholinguistic Features (LIWC)

To contextualize sentiment, we extract psycholinguistic features with LIWC. We focus on interpretable dimensions for stylistic and cultural variation, including *Affect*, *Positive Emotion*, *Negative Emotion*, *Tone*, *Social*, *Swear*, and *Conflict*. LIWC outputs are word-category proportions, which we use as normalized rates in $[0, 1]$.

4.3 Engagement Metrics

We assess how sentiment relates to engagement by correlating video-level metrics (comment counts, views, likes) with average VADER scores. We compute non-parametric rank correlations (Spearman's ρ and Kendall's τ) to account for skewed distributions. Statistical significance (p-values) is reported, but given the large sample size, interpretation focuses on effect sizes.

4.4 Polarization Metrics

We measure sentiment polarization at two levels:

4.4.1 Within-Comment Polarization

For each comment, we compute

$$\text{polarity}_{\text{comment}} = 2 \times \min(\text{pos}, \text{neg}), \tag{4.1}$$

which is high if both positive and negative proportions are substantial, and zero if sentiment is one-sided.

4.4.2 Between-Comment Polarization

For each genre, we aggregate average positive and negative proportions and define

$$\text{polarity}_{\text{genre}} = 2 \times \min(\overline{\text{pos}}, \overline{\text{neg}}). \tag{4.2}$$

High values indicate that a genre's comments contain both strong positive and strong negative sentiment across users.

4.4.3 Alternative Distribution-Based Approaches

To mitigate inflated scores from neutral-heavy distributions, we apply three distribution-based variants:

1. **Approach 1:** $1 - \text{JSD}(\text{pos}, \text{neg})$, where JSD is the Jensen–Shannon divergence.
2. **Approach 2:** A two-step filter removing low-intensity comments and excluding genres with insufficient positive/negative counts.
3. **Approach 3:** A weighted JSD incorporating pos–neg, pos–neu, and neg–neu divergences.

4.5 Correlation of LIWC and VADER

Finally, we correlate LIWC features with VADER sentiment to examine how psycholinguistic dimensions map onto lexicon-based sentiment scores. We report both coefficients and significance levels, focusing on the strength and direction of associations.

Chapter 5

Results

This chapter reports the empirical findings of the study. All procedures and metrics are defined in Section 4. We begin by analyzing how different forms of engagement relate to sentiment at the video level (comment counts, likes, views, and comment density). We then present distributions of sentiment across genres and investigate patterns of polarity and polarization. Building on this, we examine the specific case of hip-hop as an anomalous genre, before shifting focus to linguistic aspects. Finally, we analyze LIWC-based cultural and stylistic markers and connect them with VADER sentiment to contextualize communication patterns in YouTube music comments.

5.1 Engagement and Sentiment

This section examines how audience engagement relates to the sentiment expressed in comments. We distinguish between different forms of video-level engagement, including total likes, number of comments, and views, and additionally study *comment density* (comments per view) as an intensity-adjusted metric. Across all analyses, the associations are statistically significant but weak in magnitude, indicating that engagement volume influences sentiment patterns only modestly.

5.1.1 Video-level engagement

Comment count vs. sentiment

Figure 5.1 summarizes the relation between the total number of comments and average sentiment per video. Higher comment volume is associated with slightly lower positivity and higher neutral/negative tone (e.g., Spearman $\rho \approx -0.17$ for positive, $\rho \approx 0.18$ for neutral, $\rho \approx 0.15$ for negative; Kendall $\tau \approx -0.13$, 0.13, and 0.11, respectively). All reported correlations are highly significant ($p < 10^{-30}$), indicating that these relationships are unlikely to be due to random variation. While the effect sizes remain small in absolute terms, the consistent negative association with positivity and positive association with neutrality/negativity suggest that heavily discussed videos foster more neutral or critical exchanges.

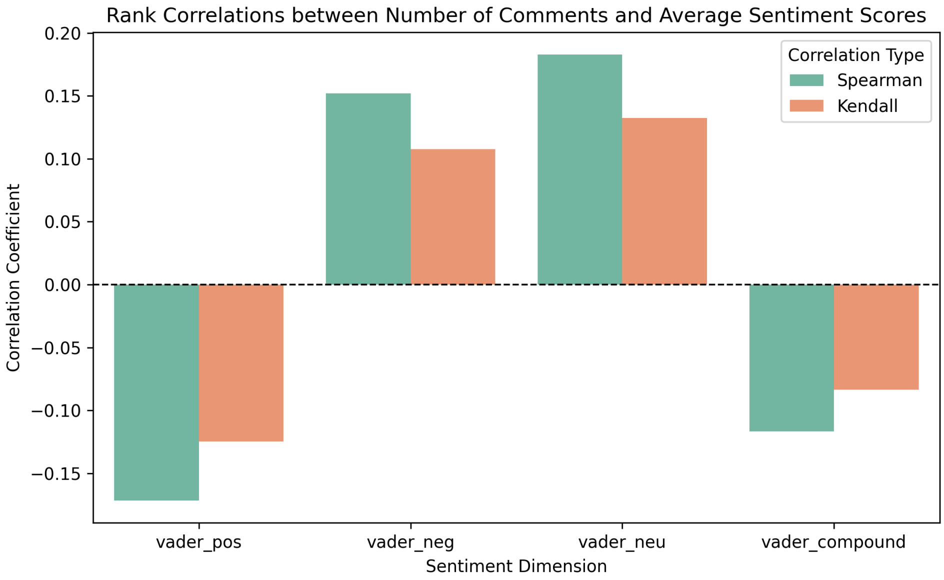

Figure 5.1: Rank correlations between number of comments and average sentiment scores (VADER) at the video level (Spearman and Kendall).

Video likes vs. sentiment

Aggregating to the video level, we correlate total likes of the Video with mean VADER scores per video. Figure 5.2 shows weak associations: videos with more likes tend to attract slightly more neutral and negative comments and slightly fewer positive ones (Spearman $\rho \approx -0.13$ for positive; $\rho \approx 0.16$ for neutral; $\rho \approx 0.12$ for negative; Kendall τ values in the range ≈ -0.10 to 0.11). All correlations are highly significant ($p < 10^{-18}$), confirming that the observed tendencies are not random. This pattern is to be expected, as videos with more likes also tend to accumulate more comments, which further strengthens the observed correlation between engagement volume and sentiment composition.

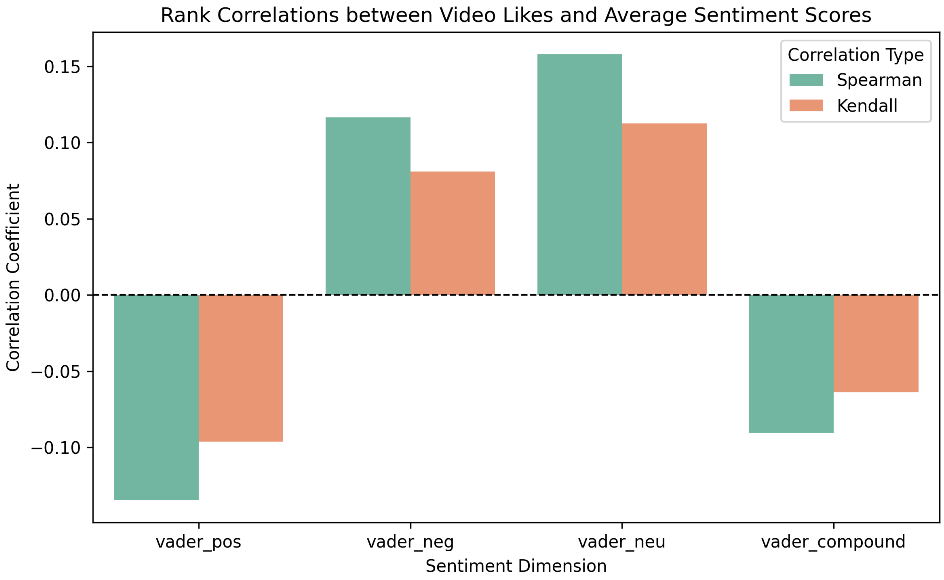

Figure 5.2: Rank correlations between video likes and average sentiment scores (VADER) at the video level (Spearman and Kendall).

Video views vs. sentiment

Correlating total `viewCount` with mean VADER scores per video reveals a similar pattern (Figure 5.3). More popular videos show slightly lower positivity and higher neutrality (Spearman $\rho \approx -0.14$ and $\rho \approx 0.17$; Kendall $\tau \approx -0.10$ and $\tau \approx 0.12$), with a smaller increase in negativity ($\rho \approx 0.09$, $\tau \approx 0.06$). All associations are highly significant ($p < 10^{-15}$), indicating consistent but weak tendencies. This confirms the patterns already observed for comment counts and likes, namely that higher engagement is linked to less positive and more neutral or critical sentiment. However, the correlations remain weak, and thus these findings should be interpreted with caution: popularity appears related to sentiment composition, but not in a strong enough way to justify definitive conclusions.

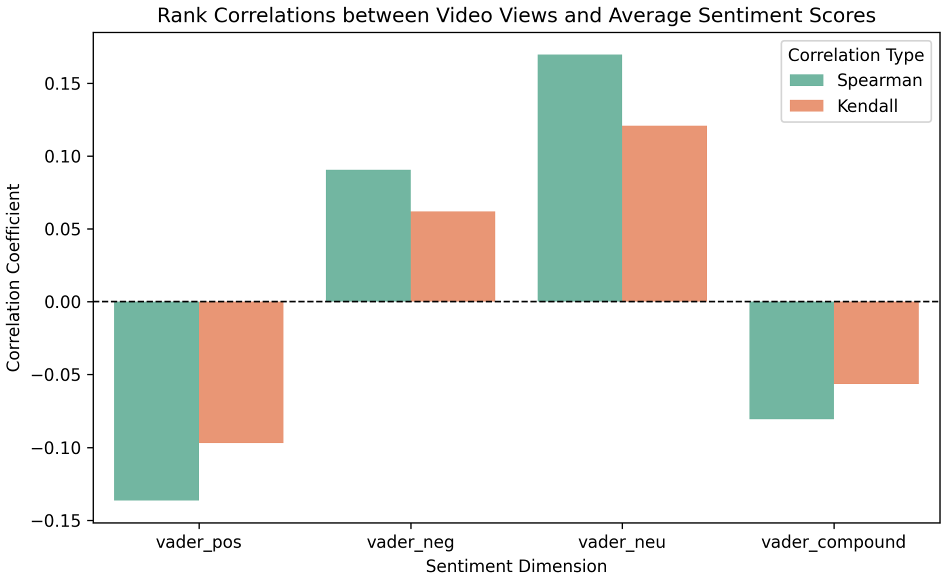

Figure 5.3: Rank correlations between video views and average sentiment scores (VADER) at the video level (Spearman and Kendall).

Comment density

To account for audience size, we analyze *comment density* (comments per view). As shown in Figure 5.4, comment density declines markedly with increasing views (Spearman $\rho = -0.43$, Kendall $\tau = -0.30$; both $p < 0.001$). This indicates that smaller, niche videos elicit proportionally more discussion, whereas mass-audience videos have lower per-view interaction.

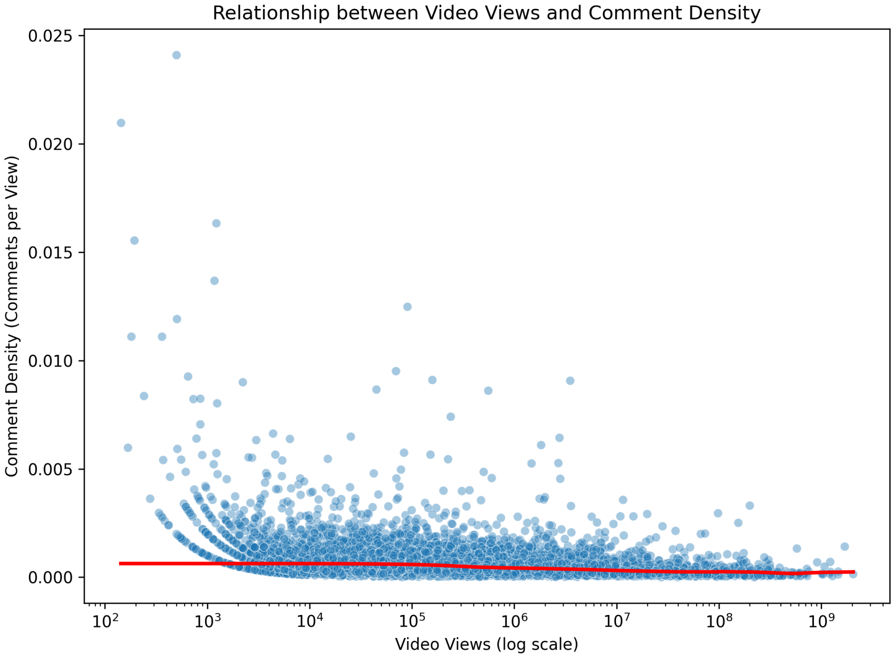

Figure 5.4: Video views (log scale) versus comment density (comments per view). A clear negative association indicates proportionally more discussion on less-viewed videos.

5.1.2 Summary of engagement effects

Across likes, views, and comment counts, higher engagement is consistently linked to slightly lower positivity and higher neutrality or negativity. Effects are statistically significant due to large n, but effect size is small ($0.1 \leq |\rho| \leq 0.18$), indicating that engagement volume only modestly shapes discussion tone. At the same time, the strong negative correlation between views and comment density shows that less popular, niche videos tend to generate proportionally more discussion than widely viewed ones.

5.2 Polarity and Polarization

We report *within-comment* polarization using Eq. (4.1) and *between-comment* polarization using Eq. (4.2). In addition, we apply the three distribution-based approaches introduced in Section 4 — (1) $1 - \mathrm{JSD}(\mathrm{pos}, \mathrm{neg})$, (2) polarity after a two-step filtering of neutral and low-intensity comments, and (3) a weighted JSD that also incorporates pos–neu and neg–neu divergences — to capture complementary perspectives on genre-level sentiment polarization.

5.2.1 Within-Comment Polarization

Extreme or intense styles (e.g., *Death Metal*, *Brutal Death Metal*, *Technical Death Metal*) show the highest within-comment ambivalence, indicating comments often mix strong positive and negative elements. *Spoken Word* and *Doom Metal* also appear among the most polarized, suggesting highly personal narratives or darker atmospheres elicit mixed emotions. By contrast, genres such as *World*, *Cool Jazz*, and *Disco* show the lowest ambivalence, with comments tending to be one-sided.

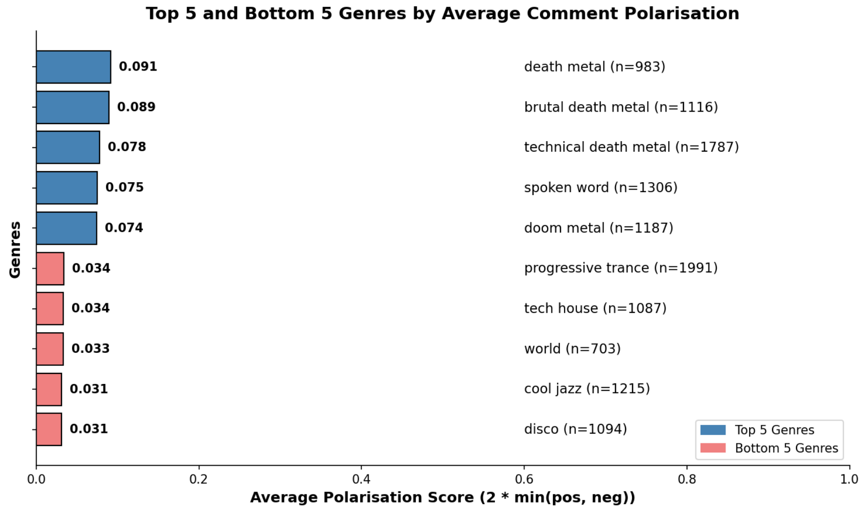

Figure 5.5: Top 5 and bottom 5 genres ranked by within-comment polarization ($2 \times \min(\text{pos}, \text{neg})$).

5.2.2 Between-Comment Polarization

Aggressive styles (*Brutal Death Metal*, *Death Metal*, *Deathcore*) are among the most polarized across comments, indicating community-level divisiveness (strong enthusiasm coexisting with strong criticism). *Noise* and *Emo Rap* also rank highly. By contrast, jazz-related genres as well as World and Progressive Trance show low polarization, indicating that comments there are more uniform in tone.

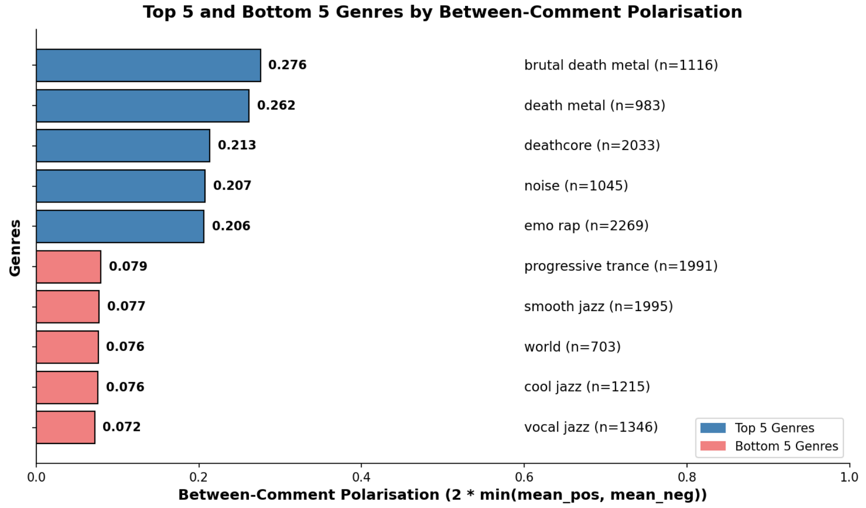

Figure 5.6: Top 5 and bottom 5 genres ranked by between-comment polarization ($2 \times \min(\overline{\text{pos}}, \overline{\text{neg}})$).

5.2.3 Alternative Approaches to Genre-Level Polarity

In addition to the within- and between-comment polarization metrics introduced in Section 4.4, we explored three alternative approaches to capture genre-level polarity more robustly. These methods aim to address potential shortcomings of the simple min-based formulation, particularly the overestimation of polarity in cases with a high proportion of neutral comments.

Approach 1: Pos–Neg Distribution Similarity. The first approach computes polarity as $1 - \mathrm{JSD}(\mathrm{pos}, \mathrm{neg})$, where JSD denotes the Jensen–Shannon divergence between the binned distributions of positive and negative sentiment scores. Intuitively, polarity is high when positive and negative distributions are similar (low divergence) and low when they diverge. This approach interprets polarity as the *balance* between positive and negative sentiment.

What the plot shows: Figure 5.7 presents the top and bottom five genres ranked by the previously described polarity definition. High-ranked genres show similar distributions of positive and negative scores, while low-ranked genres exhibit dissimilar distributions.

Interpretation:

A large majority of genres attain very high polarity scores with this approach. Out of 233 genres, 195 exceed a polarity score of 0.8. The genres with comparatively lower scores tend to be smooth, slower, and less controversial styles such as Smooth Jazz and Worship. The reason so many genres appear highly polarizing under this metric is that genres dominated by neutral comments also receive artificially high polarity values. In other words, low divergence between positive and negative distributions is not necessarily evidence of balanced debate, but can result from uniformly neutral sentiment. To address this limitation, the next approaches use filtering to reduce the inflated polarity scores caused by neutral-heavy genres.

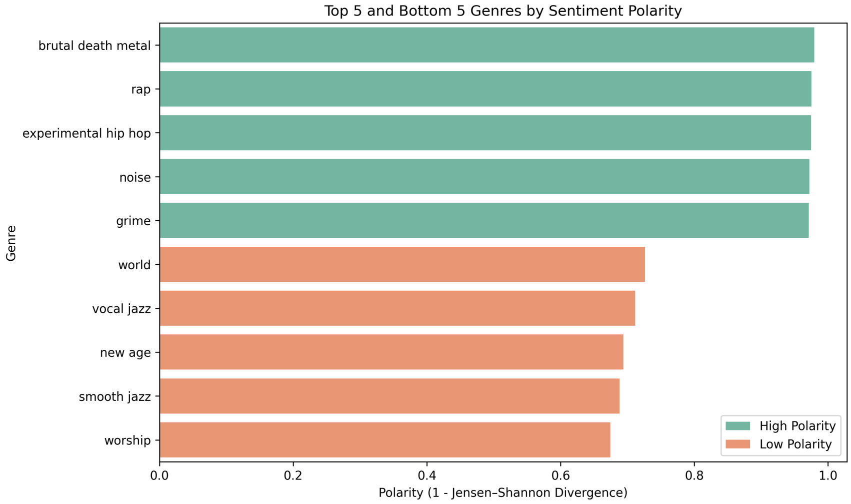

Figure 5.7: Top 5 and bottom 5 genres ranked by sentiment polarity (Approach 1: $1 - \mathrm{JSD}(\mathrm{pos}, \mathrm{neg})$). High values indicate greater similarity between positive and negative sentiment distributions.

Approach 2: Two-Step Filter.

Approach 2: Two-Step Filter. To address the inflation of polarity scores observed in Approach 1, a second refinement introduces two thresholds: (i) comments must exceed a minimum sentiment intensity (positive or negative) to be included, thus excluding largely neutral comments; (ii) each genre must contain at least a minimum number of valid positive and negative comments (e.g., 100 total comments) to ensure statistical stability. Polarity is again computed as $1 - \mathrm{JSD}(\mathrm{pos}, \mathrm{neg})$, but only after applying these filters. This approach reduces bias from neutral-heavy genres and prevents unstable estimates for genres with too few relevant comments.

What the plot shows: Figure 5.8 again shows the top and bottom five genres, but based only on genres with sufficient numbers of positive/negative comments after filtering. The plot highlights how coverage shifts once neutral-dominated or sparse genres are excluded.

Interpretation. Compared to Approach 1, the two-step filter reduced the number of included genres from 233 to 205, as genres dominated by neutral comments or with insufficient positive/negative samples were excluded. This ensures that the polarity measure is not artificially inflated by neutral-heavy distributions or statistical instability. For example, *Funk* had ranked 68th out of 233 genres in Approach 1, yet was removed entirely under this stricter filter, indicating that its apparent polarity was largely driven by a high share of neutral comments.

The relative ordering of the most polarizing genres remained stable: genres such as *Brutal Death Metal*, *Experimental Hip Hop*, and *Noise* retained top positions with only minor rank shifts ($|\Delta\mathrm{rank}| \leq 4$), although their usable comment counts were reduced by 40–50% after filtering. In contrast, many of the least polarizing genres from Approach 1 (e.g., *New Age*, *Vocal Jazz*, *World*, *Gospel*) were excluded entirely, since they lacked sufficient strong positive or negative comments. This illustrates that the filtering step not only improves robustness, but also sharpens the contrast between highly polarizing and neutral-dominated genres. Thus, the two-step filter improves robustness by emphasizing genres with substantial positive and negative engagement, while excluding those dominated by neutrality.

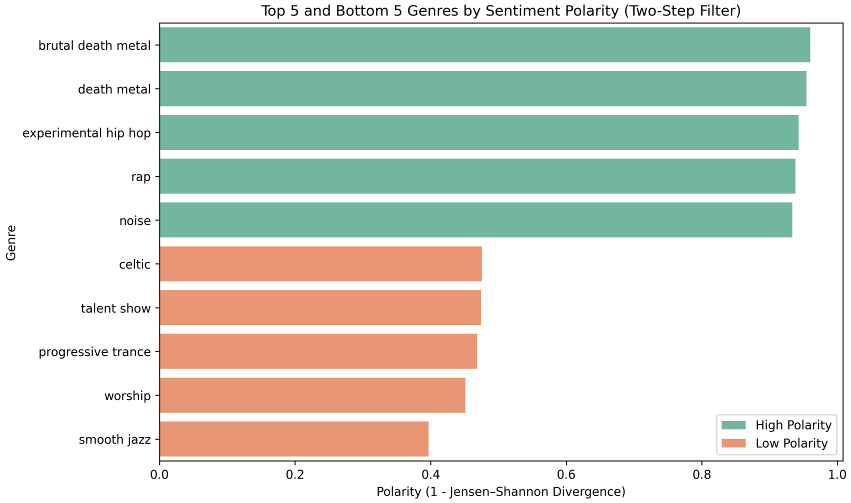

Figure 5.8: Top 5 and bottom 5 genres ranked by sentiment polarity (Approach 2: Two-step filter). Only comments exceeding a minimum sentiment threshold and genres with sufficient positive and negative comments are included.

Table 5.1: Rank changes between Approach 1 and Approach 2 for the top and bottom 10 genres. Δrank = rank_2 - rank_1 (positive = moved down, negative = moved up). Genres with fewer than 100 valid positive/negative comments after filtering were excluded.

Genre	Rank_1	Rank_2	Δrank	Comments (before → after)
Brutal Death Metal	1	1	0	1116 → 695
Rap	2	4	+2	1723 → 872
Experimental Hip Hop	3	3	0	1964 → 1042
Noise	4	5	+1	1045 → 559
Grime	5	8	+3	1587 → 773
Death Metal	6	2	-4	983 → 600
Hardcore Punk	7	6	-1	1136 → 597
Deathcore	8	7	-1	2033 → 1151
Mathcore	9	9	0	1767 → 929
Emo Rap	10	10	0	2269 → 1324
Worship	233	204	-29	1926 → 1379
Smooth Jazz	232	205	-27	1995 → 1315
New Age	231	<100	–	1102 → <100
Vocal Jazz	230	<100	–	1346 → <100
World	229	<100	–	703 → <100
Pop Rock	228	<100	–	921 → <100
Easy Listening	227	199	-28	1487 → 939
Eurovision	226	200	-26	1720 → 1122
Gospel	225	<100	–	592 → <100
Celtic	224	201	-23	1612 → 970

Approach 3: Weighted JSD of Pos/Neg/Neu. The third approach incorporates neutral comments directly by computing three pairwise Jensen–Shannon divergences:

$$\text{score} = 0.5 \cdot \text{JSD}(\text{pos}, \text{neg}) + 0.25 \cdot \text{JSD}(\text{pos}, \text{neu}) + 0.25 \cdot \text{JSD}(\text{neg}, \text{neu}),$$

and defining polarity as $1 - \text{score}$. Positive–negative divergence is weighted most strongly, but overlaps with neutrality also contribute. This reduces the risk that genres dominated by neutral comments appear maximally polarized, while still preserving information about neutrality's role in shaping sentiment.

What the plot shows: Figure 5.9 presents the top and bottom five genres under this weighted definition. Unlike Approach 1, polarity values are generally lower (all < 0.75), reflecting the penalty introduced by neutral overlap.

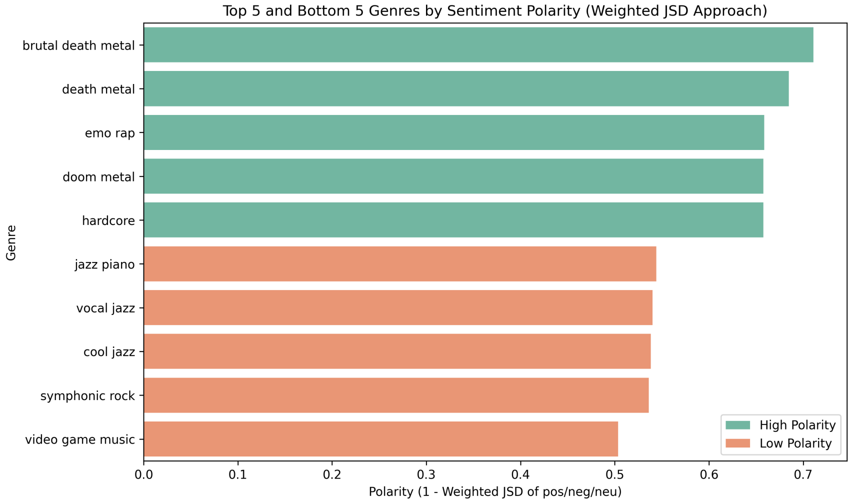

Figure 5.9: Top 5 and bottom 5 genres ranked by sentiment polarity (Approach 3: Weighted JSD of pos–neg, pos–neu, neg–neu). Neutral comments are integrated into the polarity definition with lower weights.

Interpretation. Compared to Approach 2, the weighted JSD measure introduces more pronounced rank shifts, as it explicitly penalizes overlap with neutral sentiment. While the rankings in Approaches 1 and 2 were relatively stable, several genres change their positions substantially once neutrality is incorporated.

On the top end, *Brutal Death Metal* and *Death Metal* remain firmly at the top (Ranks 1 and 2), confirming their status as highly polarized genres across all definitions. However, others such as *Experimental Hip Hop* and *Mathcore* drop sharply (Δrank $= +12$ and $+24$, respectively), reflecting the strong neutral overlap that reduces their effective polarity. Similarly, *Grime* and *Hardcore Punk* move down by 16 and 11 ranks. In contrast, *Emo Rap* improves its position significantly, moving from Rank 10 in Approach 2 up to Rank 3 here (Δrank $= -7$), indicating that it maintains strong positive–negative contrast even after neutrality is considered.

At the bottom, several genres shift markedly downward. *Blues*, *Celtic*, and *Smooth Jazz* all decline by 19–29 ranks, highlighting their heavy neutral overlap. *Worship* and *Eurovision* also move further down the scale. Interestingly, *Talent Show* improves substantially, rising 27 ranks (Rank 202 to 175), suggesting that neutrality penalization benefits certain mainstream genres that still exhibit some polarity despite many neutral comments.

Overall, this weighted approach produces a more robust scale by lowering polarity for genres with large neutral components. It sharpens the distinction between genuinely polarized genres (e.g., *Brutal Death Metal*, *Emo Rap*) and those whose apparent polarity in previous approaches was inflated by neutrality (e.g., *Mathcore*, *Blues*).

Table 5.2: Rank changes between Approach 2 and Approach 3 for the top and bottom 10 genres. $\Delta\text{rank} = \text{rank}_3 - \text{rank}_2$. Positive values indicate a genre moved down in rank, negative values that it moved up.

Genre	Rank_2	Rank_3	Δrank	Comments (App3 → App2)
Brutal Death Metal	1	1	0	1116 → 695
Death Metal	2	2	0	983 → 600
Experimental Hip Hop	3	15	+**12**	1964 → 1042
Rap	4	7	+3	1723 → 872
Noise	5	9	+4	1045 → 559
Hardcore Punk	6	17	+**11**	1136 → 597
Deathcore	7	13	+6	2033 → 1151
Grime	8	24	+**16**	1587 → 773
Mathcore	9	33	+**24**	1767 → 929
Emo Rap	10	3	**-7**	2269 → 1324
Smooth Jazz	205	224	+**19**	1995 → 1315
Worship	204	225	+**21**	1926 → 1379
Progressive Trance	203	191	-12	1991 → 1179
Talent Show	202	175	**-27**	1398 → 894
Celtic	201	228	+**27**	1612 → 970
Eurovision	200	220	+**20**	1720 → 1122
Easy Listening	199	221	+**22**	1487 → 939
Blues	198	227	+**29**	1256 → 742
Motown	197	219	+**22**	2166 → 1261
Dream Pop	196	211	+15	1487 → 854

Summary of Polarization Results. Across all approaches, polarization patterns reveal consistent tendencies: aggressive and extreme metal genres (e.g., Brutal Death Metal, Deathcore) and rap-related styles (e.g., Experimental Hip Hop, Emo Rap) emerge as the most polarized communities, while smoother or spiritual genres (e.g., Worship, Smooth Jazz, World) show the lowest polarization. The simple min-based measures (Eqs. 4.1 and 4.2) already provided a useful first indication of which genres attract audiences with more divergent opinions. The distribution-based refinements (Approaches 1–3) build on this initial intuition by incorporating intensity thresholds and explicitly accounting for neutrality, thereby yielding more robust estimates. Among these, the weighted JSD (Approach 3) offers the most balanced perspective, as it reduces artificial inflation from neutral-heavy genres while still preserving meaningful contrasts. Taken together, the results suggest that although exact rankings vary, polarization is systematically higher in genres characterized by conflictual or ambivalent expression styles, and lower in genres with more homogeneous positivity. For interpretation, the weighted JSD should be considered the most reliable indicator, as it best corrects for biases introduced by neutrality and sparse distributions.

5.3 Hip-Hop Anomaly

Hip-hop was selected for a deeper analysis because it presented a striking anomaly: despite being one of the most globally popular and culturally influential genres, it consistently ranked near the bottom in compound sentiment.

Table 5.3: Sentiment analysis for selected hip-hop genres.

Genre	**Positive Score (Rank)**	**Negative Score (Rank)**	**Neutral Score (Rank)**	**Compound Score (Rank)**
Hip Hop	0.169 (223/233)	0.080 (39/233)	0.750 (18/233)	**0.173 (215/233)**
Alternative Hip Hop	0.173 (217/233)	0.085 (23/233)	0.742 (36/233)	**0.154 (222/233)**
Experimental Hip Hop	0.156 (231/233)	0.101 (7/233)	0.743 (33/233)	**0.124 (227/233)**

Traditional hip-hop performs best among the subgenres with an average compound of 0.173 (rank 215/233), while experimental hip-hop ranks lowest at 0.124 (rank 227/233), also showing the 7th highest negative score overall.

Data Anomaly Check. One user contributed 66 of 5,418 hip-hop comments. The outlier's mean compound was -0.019; excluding the user shifts the hip-hop mean from 0.150 to 0.152 ($n = 5{,}352$), a negligible change, indicating the low scores are not driven by a single account.

Comparison with All Genres. Across 136 features, hip-hop stands out: *higher engagement* (views 16.5M vs. 10.5M, likes 188,898 vs. 82,196, comments 11,693 vs. 3,907), but *lower conventional positivity* (lower Tone and Positive Emotion), *longer comments* (20.7 vs. 17.6 words), *higher emoji usage* (6.20 vs. 4.09), and slightly *higher Authenticity*. This combination suggests that, while hip-hop attracts substantial attention, its internal communication conventions differ from those captured by lexicon-based sentiment models. The anomaly may also reflect limitations of VADER: its lexicon is not well-suited for irony, slang, or culturally specific expressions that might be common in hip-hop. Consequently, sentiment may be underestimated when such language conveys positive meaning in context but is scored as neutral or negative by the tool.

5.4 LIWC Dimensions and Cultural Markers

We compare genres on LIWC features to contextualize stylistic and cultural differences beyond sentiment.

Positive vs. Negative Balance. The balance index (`emo_pos (Positive Emotions)` − `emo_neg (negative Emotions)`) ranks spiritual and jazz genres highest, with hip-hop among the lowest, reinforcing lexicon-based lower positivity.

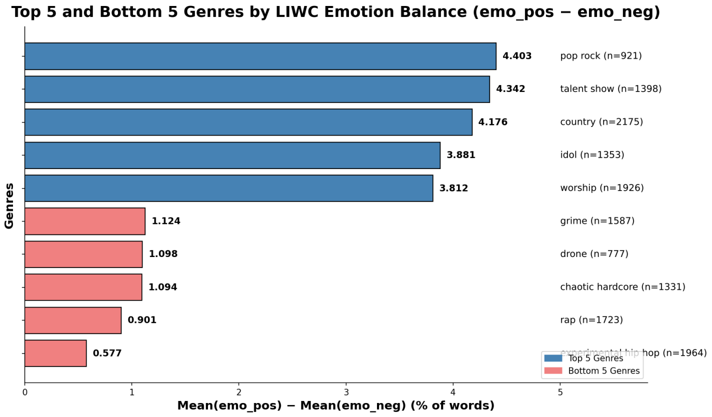

Figure 5.10: Top and bottom genres by LIWC `emo_pos` minus `emo_neg` balance.

Swear, Conflict, and Social. The `Swear` category captures the frequency of swear and offensive words, while `Conflict` refers to words denoting disagreement. Both are highest in aggressive metal genres and lowest in worship and gospel (Figures 5.11 and 5.12). In contrast, the `Social` category measures words related to social relations (e.g., friend, talk, share). This language is most frequent in *Worship*, *Country Pop*, and *K-pop*, and least common in extreme metal subgenres (Figure 5.13).

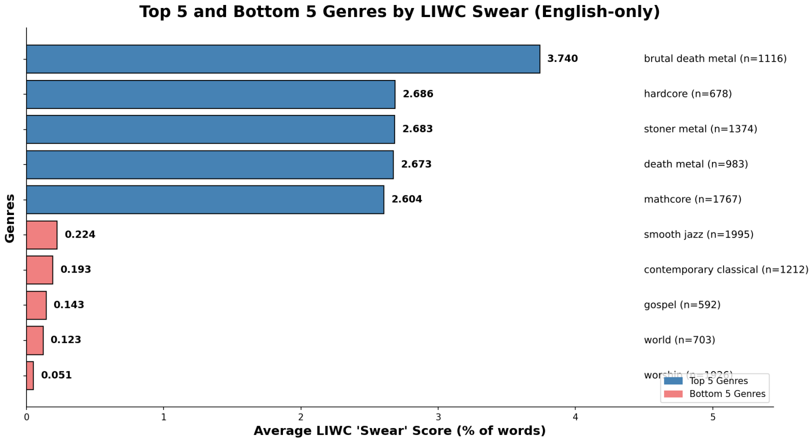

Figure 5.11: Top and bottom genres by LIWC `Swear`.

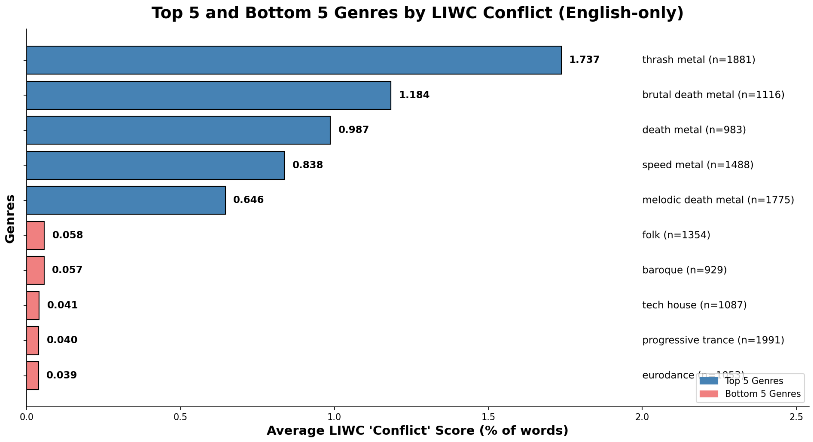

Figure 5.12: Top and bottom genres by LIWC `Conflict`.

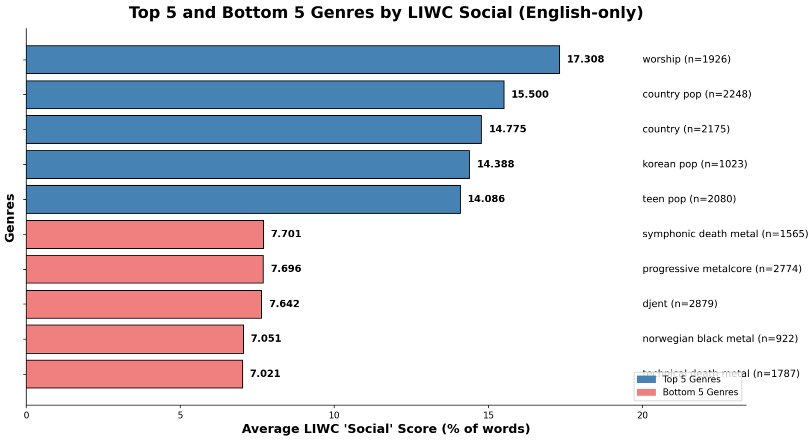

Figure 5.13: Top and bottom genres by LIWC `Social`.

5.5 LIWC–VADER Correlations

To better understand the relationship between lexicon-based sentiment analysis and psycholinguistic profiling, we correlate all LIWC dimensions with the four VADER sentiment scores (`positive`, `negative`, `neutral`, `compound`). This comparison allows us to evaluate the degree to which VADER's polarity estimates align with psychologically grounded categories such as *Affect*, *Tone*, or *Swear*, and to identify systematic correspondences as well as potential mismatches between the two tools.

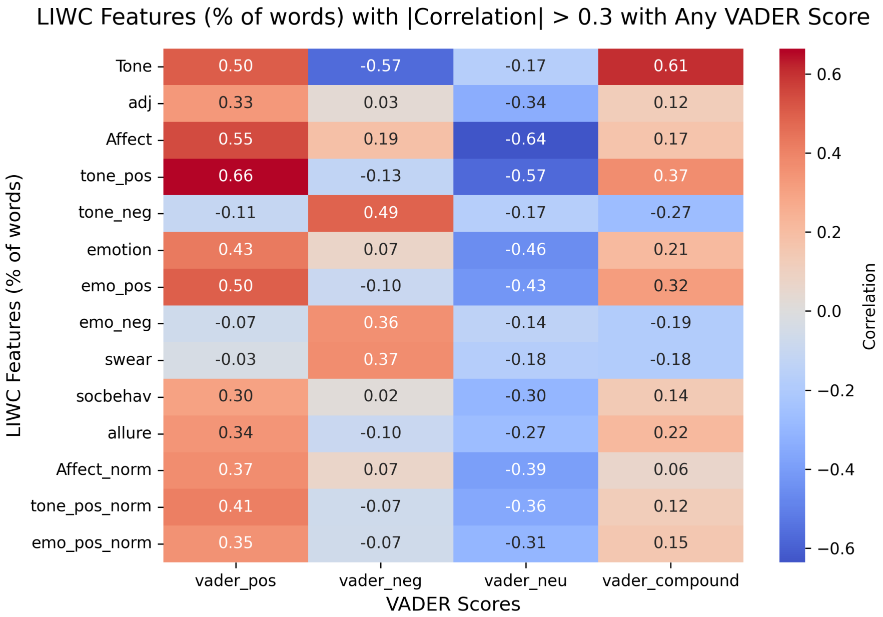

Figure 5.14: Correlation of LIWC features with VADER sentiment across all comments. Red = positive correlation; blue = negative.

Expected patterns. *Affect/Emotion/Positive Tone* show strong positive associations with VADER positive ($r \approx 0.43$–0.66). *Negative Tone/Negative Emotion* show positve correlations with VADER negative ($r \approx 0.36$–0.49). `Swear` also aligns with VADER negative ($r \approx 0.37$). As expected, the neutral score negatively correlates with *Emotion* ($r = -0.46$) and *Affect* ($r = -0.64$), indicating that comments rich in emotional language are correspondingly less likely to be classified as neutral by VADER.

Unexpected patterns. The `Emotion` category (words directly expressing feelings such as happy, love, sad, or angry) correlates much more strongly with VADER positive ($r \approx 0.43$) than with VADER negative. This is surprising, since one might expect balanced associations with both valences. A likely explanation is that the dataset is positively skewed overall, such that emotional language is more often expressed in a positive form. Another unexpected finding is that `Positive Tone` shows only a weak negative correlation with VADER negative, and `Negative Tone` shows only a weak negative correlation with VADER positive. One would expect these tone dimensions to be strongly inversely related to opposite sentiment scores.

Statistical significance. All reported correlations are statistically significant at $p < 0.001$. Given the large sample size ($n \approx 356{,}000$ comments), even weak associations reach significance, which limits the interpretive value of p-values alone. Therefore, effect size (r) is the more informative measure: moderate correlations ($r \approx 0.4$–0.6) indicate meaningful relationships, whereas weak ones ($r < 0.3$) are statistically reliable but substantively small.

Chapter 6

Conclusion and Outlook

6.1 Key Findings

This thesis set out to investigate sentiment and linguistic patterns in YouTube music video comments across genres, guided by five research questions. The main findings can be summarized as follows:

1. **Sentiment and engagement (RQ1).** Across likes, views, and comment counts, higher engagement is consistently associated with slightly lower positivity and higher neutrality or negativity. While these associations are statistically significant, effect sizes are weak ($0.1 \leq |\rho| \leq 0.18$), indicating that engagement volume only modestly shapes sentiment patterns. In contrast, niche or less popular videos generate proportionally more discussion relative to their audience size, as shown by the strong negative correlation between views and comment density. Taken together, these findings suggest that sentiment is related to engagement and popularity, but only weakly, and conclusions should therefore be drawn with caution.
2. **Genre sentiment differences (RQ2).** Clear cross-genre differences emerged. Spiritual and smooth genres (e.g., worship, gospel, smooth jazz) showed the most positive average sentiment. Aggressive (e.g., death metal, noise, hardcore punk) scored lowest, yet still above zero, indicating that even the "most negative" genres lean slightly positive overall.
3. **Polarization (RQ3).** Both within-comment and between-comment analyses reveal clear genre-specific polarization patterns. Aggressive and rap-related styles (e.g., Brutal Death Metal, Deathcore, Experimental Hip Hop, Emo Rap) consistently show high polarization, indicating that these communities combine strong enthusiasm with equally strong criticism. By contrast, smoother and spiritual genres (e.g., Worship, Smooth Jazz, World) exhibit low polarization, with comments being more uniformly positive. The distribution-based refinements (Approaches 1–3) further demonstrate that filtering for sentiment intensity and accounting for neutrality improves robustness, with the weighted JSD (Approach 3) providing the most balanced view. Taken together, these findings suggest that polarization is systematically higher in communities where conflictual or ambivalent expression is part of the communication style, and lower in communities characterized by homogeneous positivity.
4. **Psycholinguistic features (RQ4).** LIWC analysis highlights clear genre-specific linguistic markers. Aggressive metal genres contain the highest frequencies of swearing and conflict terms, aligning with their lower VADER positivity and higher polarization. By contrast, worship and gospel genres are characterized by high levels of social and religious language, corresponding to more uniformly positive sentiment. These results demonstrate that psycholinguistic features not only distinguish genres

but also interact with sentiment: categories such as `Swear` and `Conflict` are positively correlated with negative sentiment, while `Social` and `Positive Emotion` tend to align with positivity. This confirms that linguistic style provides important context for interpreting sentiment patterns across music communities.

Taken together, the results confirm that genre context matters greatly for interpreting online sentiment, and that lexicon-based methods should be applied with caution when analyzing diverse cultural communities.

6.2 Limitations

While the analyses provide novel insights, several limitations must be acknowledged:

- **Language filtering:** Only English comments were retained, excluding large non-English communities and possibly biasing results for globally popular genres.
- **Tool limitations:** VADER and LIWC cannot fully capture irony, sarcasm, or the positive use of swearing and slang. This limitation could be particularly relevant for hip-hop, which may help explain the observed anomaly.
- **Temporal scope:** Restricting to the 100 most recent comments per video reflects current mood but not historical sentiment trends.
- **Genre coverage:** Genres with fewer than 500 comments were excluded, omitting smaller or emerging communities.
- **Engagement measures:** Likes and views are shaped by YouTube's algorithms and trends, and may not directly represent audience sentiment.

6.3 Outlook

Future research can build on this work in several ways:

- **Multilingual extension:** Including non-English comments would allow analysis of global communities and cross-cultural comparisons.
- **Advanced sentiment models:** Transformer-based or fine-tuned large language models could better capture slang, sarcasm, and cultural nuance than lexicon-based tools.
- **Temporal dynamics:** Tracking comment sentiment over time could reveal how community mood shifts around events like album releases or controversies.
- **Broader engagement measures:** Going beyond likes/views to include replies, shares, or watch time would provide a more complete picture of audience interaction.

Overall, this thesis provides the first large-scale, genre-aware analysis of sentiment and linguistic markers in YouTube music video comments. It demonstrates both expected and surprising patterns, and lays the groundwork for more culturally and methodologically robust future research.

Bibliography

[1] Rauf Arif, Yoonhyeung Cho, et al. 2024. Displays of Parasocial Interaction in K-Pop: A Content Analysis of YouTube Comments on BTS's Music Videos. *Nordicom Review*, 45, 1, 67–86. DOI: 10.2478/nor-2024-0005. https://www.researchgate.net/publication/382348670_Displays_of_Parasocial_Interaction_in_K-Pop_A_Content_Analysis_of_YouTube_Comments_on_BTS%27s_Music_Videos.

[2] Christine Bauer and Markus Schedl. 2019. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. In *Proceedings of the 17th IFIP International Conference on Human-Computer Interaction (INTERACT 2019)*. Springer, pp. 589–609. DOI: 10.1007/978-3-030-29387-1_35. https://doi.org/10.1007/978-3-030-29387-1_35.

[3] Seung Woo Chae, Noriko Hara, Harshit Rakesh Shiroiya, Janice Chen, and Ellen Ogihara. 2024. Being Vulnerable with Viewers: Exploring how medical YouTubers communicated about COVID-19 with the public. *PLOS ONE*, 19, 12, e0313857. DOI: 10.1371/journal.pone.0313857. https://doi.org/10.1371/journal.pone.0313857.

[4] Ioannis Chalkias, Katerina Tzafilkou, Panagiotis Karapiperis, and Christos Tjortjis. 2023. Exploring Sentiment Analysis Methods and Topic Clustering. *Electronics*, 12, 18, 3949. DOI: 10.3390/electronics12183949. https://www.mdpi.com/2079-9292/12/18/3949.

[5] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. https://github.com/microsoft/ToxiGen.

[6] Hausberger, Strauß, and Schedl. 2025. ExIM: Exploring Intent of Music Listening for Retrieving User-Generated Playlists. Preprint. (2025).

[7] Ricardo Hernández-Pérez, Pablo Lara-Martínez, Bibiana Obregón-Quintana, Larry S. Liebovitch, and Lev Guzmán-Vargas. 2024. Correlations and Fractality in Sentence-Level Sentiment Analysis Based on VADER for Literary Texts. *Information*, 15, 11, (November 2024), 698. DOI: 10.3390/info15110698. https://www.mdpi.com/2078-2489/15/11/698.

[8] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8, 1, (June 2014), 216–225. DOI: 10.1609/icwsm.v8i1.14550. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[9] Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. Music4All-Onion. Zenodo. https://zenodo.org/records/6609677.

[10] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. Technical report. University of Texas at Austin.

[11] Rahul Singha. 2024. Spam Detection on YouTube Comments Using Advanced Machine Learning Models: A Comparative Study. *arXiv preprint arXiv:2401.12345*. Accessed: 2024-07-31.

[12] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29, 1, 24–54.

[13] 2024. YouTube: number of monthly active users worldwide as of January 2024. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed: 2024-07-31. (2024).

[14] Haowen Zhang. 2025. Sentiment Analysis on YouTube Data: A Comparison of TextBlob and VADER. *Frontiers in Humanities and Social Sciences*, 5, 1, 71–78. DOI: 10.26855/fhss.2025.01.010. https://www.hillpublisher.com/ArticleDetails/4437.

Appendix A

Representative Song Selection Algorithm

The algorithm operates in an iterative fashion to ensure fair distribution of representative videos across genres.

Input:

- For each genre: ranked list of candidate videos with relevance scores.
- Parameter k: number of representative videos per genre.

Output:

- For each genre: exactly k unique videos.

Procedure:

1. **Initial Assignment:** Assign the top k videos to each genre with at least k candidates.
2. **Conflict Detection:** Identify videos assigned to multiple genres.
3. **Conflict Resolution:**
   - Compute fallback scores for each genre (the score of the next unused candidate).
   - Retain the video for the genre with the lowest fallback score (i.e., the worst alternative).
   - Break ties using: (i) higher score on contested video, (ii) alphabetical order of genre name.
4. **Loser Refill:** Genres losing a conflict replace the video with their next best unused candidate.
5. **Iteration:** Repeat conflict detection and resolution until no conflicts remain or candidates are exhausted.
6. **Finalization:** Remove genres with fewer than k assigned videos and output the final assignment.