computers

MDPI

Article

Cross-Linguistic Complexity and Language-Specific Sentiment: Multifractal Structure and Emotional Valence in Popular Music Lyrics Across Three Languages

Fateme Khanipour [1], Zeinab Shahbazi [2,*], Sara Behnamian [3,*], Fatemeh Fogh [4] and Nathan Blood [4]

[1] Independent Researcher, Tehran 1477636718, Iran; fateme.khanipour@gmail.com
[2] Research Environment of Computer Science (RECS), Kristianstad University, 291 88 Kristianstad, Sweden
[3] Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark
[4] Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA; ffogh2021@fau.edu (F.F.); nblood2022@fau.edu (N.B.)
* Correspondence: zeinab.shahbazi@hkr.se (Z.S.); sara.behnamian@sund.ku.dk (S.B.)

Abstract

We investigate the linguistic complexity and emotional valence of popular song lyrics across English ($n = 1491$), Spanish ($n = 307$), and German ($n = 225$), using an analytical corpus of 2023 tracks drawn from 2113 deduplicated tracks on Spotify's weekly Top 200 charts (2019–2021). Transformer-based sentiment analysis is combined with complexity-science tools to characterize both the affective content and the structural organization of commercially successful lyrics. A multilingual BERT model reveals a mild negative skew across all three languages (63.7% negative overall); the 1.003-point English–German gap observed under the English-centric VADER lexicon collapses to 0.127 points under BERT, indicating that earlier cross-linguistic sentiment differences are largely measurement artifacts. Word frequency distributions follow Zipf's law in all three languages ($R^2 > 0.96$), with English steepest ($\alpha = 1.409$) and German shallowest ($\alpha = 1.181$). Detrended fluctuation analysis indicates persistent long-range correlations ($H \approx 0.66$–0.76; none of the 50 shuffled surrogates exceeded the observed values), and multifractal singularity spectra are statistically indistinguishable across languages once corpus size is controlled (all pairwise Mann–Whitney $p > 0.13$). Streaming counts within the Top 200 are concentrated (German Gini $= 0.556$) but, given the truncated single-snapshot sample, are reported as within-chart descriptors rather than population-level scaling.

Keywords: song lyrics; sentiment analysis; BERT; Zipf's law; multifractal analysis; detrended fluctuation analysis; Spotify; cross-linguistic; complexity science; natural language processing

Check for updates

Academic Editors: Phivos Mylonas, Christos Troussas, Akrivi Krouska, Manolis Wallace and Maria Kouri

Received: 10 April 2026
Revised: 9 May 2026
Accepted: 12 May 2026
Published: 14 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Popular music lyrics occupy a unique position at the intersection of linguistic creativity and mass culture. Unlike many other written genres, commercially successful song lyrics must simultaneously satisfy aesthetic, affective, and formal constraints, conveying emotional content within the phonological, rhythmic, and structural demands imposed by musical composition. These competing pressures make lyrics a particularly rich corpus for examining the statistical and linguistic regularities that emerge at the boundary between constrained and expressive language use [1,2]. Yet despite the global reach of streaming platforms and the extraordinary scale of the data they generate, quantitative linguistic

study of popular lyrics has remained largely fragmented along disciplinary lines: music information retrieval (MIR), computational linguistics, and the physics of complex systems have each developed analytical traditions that rarely speak to one another [3–5].

The computational analysis of sentiment and emotion in song lyrics has a substantial history, much of it focused on automatic mood classification within music information retrieval and recommendation systems [6,7]. Early approaches relied on bag-of-words representations and domain-specific lexicons, with VADER (Valence Aware Dictionary and sEntiment Reasoner; [8]) becoming a widely used baseline due to its strong performance on short, informal social-media text. However, VADER was constructed exclusively from English-language data, meaning that its application to non-English lyrics is methodologically suspect: the lexicon fails to recognize sentiment-bearing vocabulary in other languages, typically producing artificially negative scores when tokens are simply unlisted rather than genuinely negative. This limitation is rarely acknowledged in the music sentiment literature, which has largely conducted analyses within single-language corpora [1,3], often without explicitly addressing potential cross-linguistic measurement artifacts.

The development of multilingual transformer models, particularly multilingual BERT variants that can be fine-tuned for sentiment classification, has made it possible to revisit cross-linguistic comparisons with considerably greater methodological rigor [9]. Pre-trained on text from over 100 languages and fine-tuned for star-rating prediction on product reviews spanning English, Dutch, German, French, Italian, and Spanish, multilingual BERT [10,11] can produce sentiment estimates that are genuinely comparable across languages, rather than reflecting the coverage asymmetries of English-centric lexicons. An open empirical question is how much of the cross-linguistic sentiment variation reported in prior work survives when an equitable multilingual model is applied, with direct implications for theories linking language typology to the emotional character of popular music.

In parallel, a largely independent line of research has applied tools from statistical physics to the analysis of natural language structure. Zipf's law—the observation that word frequency is approximately inversely proportional to word rank—has been documented extensively across multiple languages [12–14], and its behavior in constrained registers such as song lyrics has received growing but still limited attention [15]. At the same time, detrended fluctuation analysis (DFA; [16]) and its multifractal extension (MF-DFA; [17]) have revealed persistent long-range correlations in literary texts [18–20], indicating that natural language carries structured temporal memory at multiple scales [4]. Whether these properties extend to the highly constrained and repetitive register of popular song lyrics remains unexplored.

The present study brings these three strands of inquiry together in an integrated cross-linguistic analysis of Spotify chart lyrics across English, Spanish, and German. Using a corpus of 2023 unique tracks drawn from 54 weekly Top 200 chart snapshots (2019–2021), we address four interrelated questions: (1) Do multilingual BERT-based sentiment models produce cross-linguistically consistent estimates compared to the English-centric VADER lexicon [8]? (2) Do lyrics word frequency distributions adhere to Zipf's law [12,13,21], and do they vary in slope and lexical richness across languages? (3) Do lyrics exhibit long-range correlations and multifractal scaling [16–18], and are these properties shared across the three languages examined or language-specific? (4) How concentrated is the within-Top-200 streaming distribution, and does this within-chart concentration vary across English, Spanish, and German language markets?

By combining transformer-based sentiment analysis with complexity science methods in a single large-scale cross-linguistic corpus, this study provides a methodologically rigorous account of the statistical properties of commercially successful song lyrics, with im-

plications for computational musicology, cross-linguistic NLP, and the broader study of language use under formal compositional constraints.

2. Literature Review

2.1. Sentiment Analysis in Music Lyrics

The automatic analysis of sentiment and mood in song lyrics has developed substantially over the past two decades, driven largely by the demands of music recommendation and information retrieval systems. Early work established that lyrics carry meaningful affective signals that complement audio-based features. Laurier et al. [7] demonstrated that combining lyrics and audio information can improve mood classification performance by leveraging complementary information from both modalities. Similarly, Hu and Downie [3] showed that lyric-based features contribute significantly to mood classification in music digital libraries, and that integrating lyrics with audio features outperforms systems relying solely on audio in experiments conducted on a dataset covering 18 mood categories.

The breadth of approaches for representing lyrical affect was surveyed by Kim et al. [6], who provided a comprehensive review of music emotion recognition research, summarizing psychological models of musical emotion as well as computational approaches based on audio content, contextual textual information (such as lyrics and tags), and multimodal systems. Lexicon-based approaches have remained influential due to their interpretability and computational simplicity.

Hutto and Gilbert [8] developed VADER, a rule-based sentiment analyzer optimized for short, informal text that became a widely adopted baseline for lyrics-level sentiment scoring. However, VADER's construction from exclusively English-language data severely limits its cross-linguistic applicability: when applied to non-English lyrics, the lexicon produces systematically biased scores that reflect missing vocabulary rather than genuine sentiment differences. The rise of transformer-based models has provided an avenue for addressing this limitation. Devlin et al. [9] introduced BERT and its multilingual variant, which encode contextual representations across over 100 languages and can be fine-tuned for sentiment classification in a linguistically equitable manner.

The computational study of lyrics has also been enriched by broader work in lyrics-based classification: Fell and Sporleder [1] introduced an approach for automatic genre classification from lyrical features, demonstrating that linguistic properties of lyrics—vocabulary, style, semantics, and structural orientation—carry robust predictive signals well beyond simple keyword matching.

2.2. Zipf's Law and Lexical Statistics in Language and Music

The power-law relationship between word rank and frequency, first formalized by Zipf [12], is widely regarded as one of the most robust empirical regularities in quantitative linguistics. Piantadosi [13] provided a thorough critical review of theoretical accounts of Zipf's law, documenting its cross-linguistic universality while noting systematic variation in exponents across languages and registers. Corral et al. [21] further investigated Zipf's law by comparing word-form and lemma representations in long literary texts across several languages with varying levels of morphological complexity. Their analysis showed that power-law frequency distributions hold for both word forms and lemmas over several orders of magnitude, while the estimated Zipf exponents remain broadly similar despite the substantial transformation introduced by lemmatization.

The application of Zipf's law beyond natural language text to musical corpora has received growing attention. Perotti et al. [22] demonstrated that Zipf's law emerges in musical compositions when appropriate Zipfian units—combinations of notes and chords—are identified, suggesting a deep structural analogy between music and language.

Williams et al. [15] further showed that Zipfian scaling extends to phrase-level rather than purely word-level units, with power-law distributions holding over substantially more orders of magnitude when phrases are treated as the fundamental unit of analysis. Together, these findings motivate the investigation of Zipf's law in the constrained lexical register of popular song lyrics, where vocabulary is systematically restricted relative to literary prose.

2.3. Long-Range Correlations and Multifractal Scaling in Text

The application of methods from statistical physics to textual corpora has revealed that natural language is far from a random sequence of words. Peng et al. [16] introduced detrended fluctuation analysis (DFA), a technique for quantifying long-range correlations in non-stationary time series, originally developed for DNA sequence analysis. The method was subsequently adapted for linguistic time series by representing texts as sequential signals of word length or frequency rank and estimating the Hurst exponent H, where $H > 0.5$ indicates persistent long-range dependence. Kosmidis et al. [4] applied DFA to word-frequency time series in literary texts, confirming the presence of significant long-range correlations and demonstrating that the structure of language cannot be captured by short-memory models alone. Altmann et al. [23] complemented this picture by showing that word usage in online corpora is shaped by niche-like constraints, where the dissemination of words across users and topics strongly influences their temporal dynamics and likelihood of persistence.

The multifractal extension of DFA, introduced by Kantelhardt et al. [17], provides a method for characterizing the multifractal scaling behavior of nonstationary time series through a spectrum of generalized Hurst exponents. The method also makes it possible to distinguish multifractality arising from long-range correlations from that associated with broad probability distributions.

Montemurro and Pury [18] provided foundational evidence for long-range fractal correlations in literary texts, showing that sequences of word ranks exhibit persistent scaling behavior characterized by Hurst exponents significantly larger than 0.5. Grabska-Gradzińska et al. [20] extended this to sentence-length series in English literary texts using both MF-DFA and wavelet-based methods, identifying a subset of texts with genuine multifractal structure distinguishable from monofractal or structureless controls. Most comprehensively, Drożdż et al. [19] quantified the origin and character of long-range correlations in a large corpus of narrative texts by analyzing sentence-length variability. Using power-spectral analysis, multifractal detrended fluctuation analysis, and surrogate data tests, they showed that many texts exhibit scale-free correlations and that nonlinear correlations contribute to the multifractal organization of sentence structures.

2.4. Power-Law Scaling in Streaming and Popularity Distributions

Beyond the internal structure of lyrics, the distributional properties of streaming activity raise independent empirical questions. Online content diffusion and popularity distributions have been extensively studied in the context of social platforms. Goel et al. [24] analyzed the structure of online diffusion networks, finding that popularity distributions are heavily right-skewed and consistent with power-law or log-normal scaling, driven by preferential attachment dynamics in which early popularity advantages compound over time. While such population-level scaling cannot be tested with chart-only data of the kind used in the present study, these findings motivate the examination of how concentrated the within-chart streaming distribution of chart songs is, and whether this within-chart concentration—quantified by the Gini coefficient—varies systematically across language markets.

Table 1 summarizes key related work on lyrics sentiment analysis and textual complexity, highlighting the gap that the present study addresses by combining sentiment analysis with complexity methods in a cross-linguistic lyrics corpus.

Table 1. Summary of related work on lyrics sentiment analysis and textual complexity.

Study	Domain	Language(s)	Corpus Size	Sentiment Method	Complexity Methods	Cross-Ling.
Fell & Sporleder [1]	Song lyrics	English	Not reported	Lexicon features	—	No
Hu & Downie [3]	Song lyrics	English	Not reported	Bag-of-words + audio	—	No
Napier & Shamir [25]	Song lyrics	English	∼6000 songs	Lexicon-based	—	No
Hong et al. [26]	Song lyrics	English, Chinese	Not reported	BERT, CNN, BiLSTM	—	Yes
Kosmidis et al. [4]	Literary texts	English, Greek	36,221 words	—	DFA, GP	Yes
Montemurro & Pury [18]	Literary texts	English	Multiple texts	—	DFA	No
Grabska-Gradzińska et al. [20]	Literary texts	English	∼100 texts	—	MF-DFA, WTMM	No
Drożdż et al. [19]	Literary texts	English	Large corpus	—	MF-DFA, PSD	No
De Santis et al. [27]	Literary texts	5 language families	Multiple corpora	—	MF-DFA	Yes
Tanaka-Ishii & Takahashi [28]	Literary texts	10 languages	Multiple corpora	—	Taylor, EN methods	Yes
Gillet & Ausloos [29]	Literary texts	English, Esperanto	2 texts	—	MF analysis	Yes
Present study	Song lyrics	English, Spanish, German	2023 songs	VADER + multilin. BERT	Zipf, DFA, MF-DFA	Yes

Note. DFA = detrended fluctuation analysis; MF-DFA = multifractal DFA; GP = Grassberger–Procaccia; PSD = power spectral density; WTMM = wavelet transform modulus maxima; EN = Ebeling & Neiman.

3. Materials and Methods

Figure 1 provides an overview of the complete analytical pipeline, from data collection through sentiment and complexity analyses.

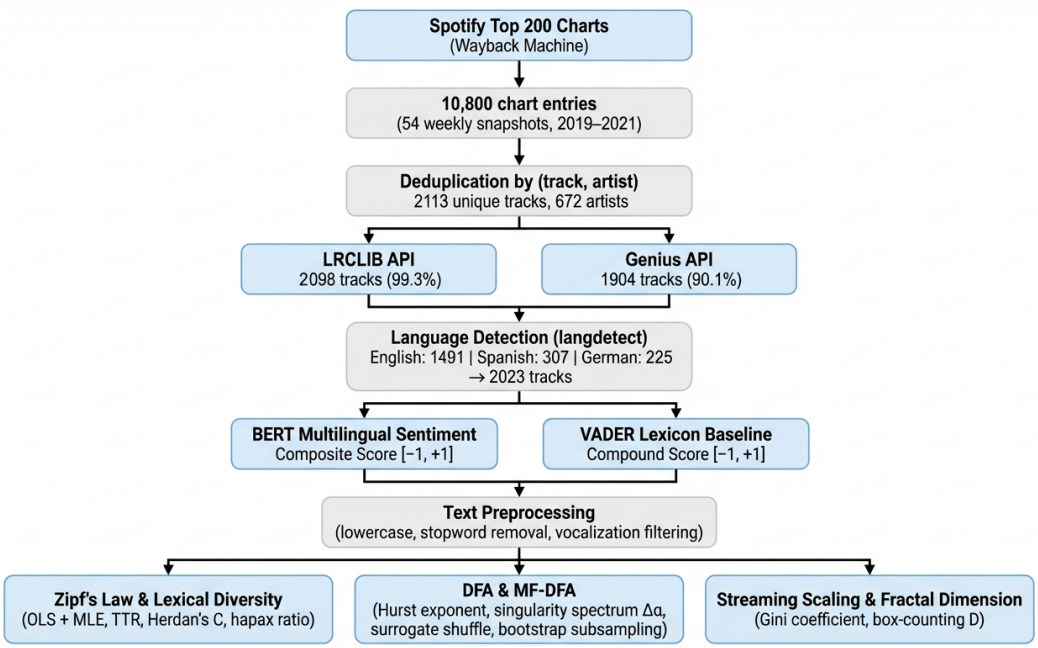


Figure 1. Overview of the analytical pipeline. Chart data were collected from Spotify Top 200 weekly snapshots via the Wayback Machine and deduplicated to 2113 unique tracks. Lyrics were retrieved from two independent sources: LRCLIB, used for BERT multilingual sentiment analysis and all downstream complexity analyses, and Genius, used for the VADER lexicon-based sentiment baseline. Language detection restricted the corpus to English, Spanish, and German ($N = 2023$). Arrows indicate the sequential flow of data processing and analysis, and colors distinguish the main analytical stages. Text preprocessing was applied prior to Zipf's law, multifractal, and streaming distribution analyses but not to BERT sentiment scoring (see Section 3.6).

3.1. Data Collection

Chart performance data were obtained from Spotify's historical weekly Top 200 charts via the Wayback Machine internet archive (web.archive.org). The Common Data Index

(CDX) API was queried for archived snapshots of the spotifycharts.com download endpoints across seven regional markets: Global, United States, United Kingdom, Germany, France, Spain, and Poland. The collection targeted the period 2019–2023; however, no usable archival snapshots of the spotifycharts.com download endpoints were retrieved beyond 2021. The retrieved dataset therefore effectively spans 2019–2021 and comprises 54 weekly chart snapshots distributed unevenly across three years: 2019 ($n = 31$ snapshots, 6200 chart entries), 2020 ($n = 17$ snapshots, 3400 entries), and 2021 ($n = 6$ snapshots, 1200 entries), yielding 10,800 chart entries in total. All subsequent references to the temporal coverage of the corpus refer to this 2019–2021 effective window. Each entry contained the track name, artist name, chart position (1–200), and weekly stream count.

To construct a corpus of unique songs, duplicate entries arising from tracks appearing across multiple weekly snapshots were removed on the basis of the (track name, artist name) combination, retaining the first occurrence. This yielded 2113 unique tracks by 672 distinct artists, which constitute the deduplicated full corpus. The analytical subset of 2023 tracks used in all sentiment and complexity analyses corresponds to tracks classified as English, Spanish, or German by the language detection step (Section 3.3). The remaining 90 tracks comprised songs in other detected languages or tracks classified as unknown because lyrics were not retrieved. Throughout the remainder of this paper, "deduplicated full corpus" refers to the 2113 tracks and "analytical corpus" to the 2023 English, Spanish, and German tracks. The retained stream count for each track reflects a single weekly snapshot rather than a cumulative, peak, or average popularity measure; stream counts in the deduplicated full corpus ranged from 284,556 to 65,873,080 ($M = 6{,}869{,}596$; $SD = 5{,}937{,}902$; $Mdn = 5{,}567{,}732$).

3.2. Lyrics Retrieval

Lyrics were retrieved from two independent sources to support both lexicon-based and transformer-based sentiment analyses. The primary source was LRCLIB (lrclib.net), a community-maintained lyrics database accessed via its public REST API without authentication. For each of the 2113 unique tracks, a search query was issued using track name and artist name. LRCLIB returned matching lyrics for 2098 tracks (99.29% of the deduplicated full corpus), while 6 tracks were identified as instrumental and 9 were not found. Median lyrics length was 2047 characters ($M = 2183.8$; $SD = 887.2$; range: 46–7772 characters). Only one track had lyrics shorter than 100 characters.

As a secondary source, the Genius API (genius.com) was queried independently for the same 2113 tracks. Genius returned lyrics for 1904 tracks (90.11%), with 209 tracks unmatched. The Genius-sourced lyrics served as the basis for a lexicon-based sentiment baseline using the VADER sentiment analyzer [8], while the LRCLIB-sourced lyrics were reserved for transformer-based sentiment analysis described in Section 3.6. Rate limits of 0.3 s per request (LRCLIB) and 0.5 s (Genius) were imposed to comply with API usage policies.

3.3. Language Detection

Language identification was performed on the LRCLIB-sourced lyrics using the `langdetect` Python library, version 1.0.9 [30], which implements a Naïve Bayes classifier trained on Wikipedia character n-gram profiles. Prior to detection, lyrics were preprocessed by removing common non-lexical vocalizations (e.g., "uh," "yeah," "oh no"), non-alphabetic characters, and excess whitespace. The detector's random seed was fixed (seed = 0) to ensure reproducibility.

A total of 18 distinct languages were identified across the corpus. English was the dominant language ($n = 1491$; 70.56%), followed by Spanish ($n = 307$; 14.53%) and German ($n = 225$; 10.65%). These three languages jointly accounted for 2023 tracks (95.74%

of the deduplicated full corpus). The remaining 90 tracks comprised 75 tracks assigned to 15 additional detected languages, each represented by fewer than 30 tracks, and 15 tracks (0.71%) classified as "unknown" because no lyrics were retrieved.

3.4. Baseline Sentiment Analysis (VADER)

A lexicon-based sentiment baseline was established using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analyzer [8], applied to Genius-sourced lyrics. VADER computes four scores for each text: a positive proportion (pos), negative proportion (neg), neutral proportion (neu), and a normalized compound score ranging from -1.0 (most negative) to $+1.0$ (most positive). Following established practice in music sentiment research, songs were classified as positive if pos $>$ neg, and negative otherwise. This baseline provides a reference against which the transformer-based BERT analysis (Section 3.6) is compared.

3.5. Data Integration and Language Filtering

The VADER sentiment scores derived from Genius lyrics (Section 3.4) were integrated with the LRCLIB-based lyrics corpus (Section 3.2) via a two-stage encoding-normalized merge on (track name, artist name). The first stage applied Unicode NFKD decomposition with diacritical mark removal to both datasets; the second stage stripped all non-alphanumeric characters to resolve residual mismatches arising from character encoding discrepancies between data sources (e.g., "ROSALÍA" vs. "ROSAL?A"). This procedure matched 1833 of the 1904 Genius-scored tracks (96.27%); the remaining 71 tracks could not be linked due to track-level mismatches across sources.

For all subsequent analyses, the corpus was restricted to the three dominant languages: English ($n = 1491$), Spanish ($n = 307$), and German ($n = 225$), yielding an analytical subset of 2023 tracks (95.74% of the deduplicated full corpus). Of these, 1771 had VADER sentiment scores (English: 1312; Spanish: 246; German: 213).

3.6. Multilingual BERT Sentiment Scoring

To address the inherent limitations of applying the English-centric VADER lexicon to multilingual lyrics, a transformer-based sentiment analysis was conducted using the `nlptown/bert-base-multilingual-uncased-sentiment` model [9]. This model, fine-tuned for multilingual sentiment classification on product reviews across six languages (including English, Spanish, and German), outputs a probability distribution over five sentiment classes (1-star to 5-star). Sentiment was scored on the original, unprocessed lyrics rather than the stopword-removed text used in downstream lexical analyses, as BERT's contextual attention mechanism relies on grammatical structure and function words for accurate inference [10].

Two continuous sentiment indices were derived from the five-class probability vector. The BERT Composite Score was computed as the dot product of the probability vector with the weight vector $[-1.0, -0.5, 0.0, +0.5, +1.0]$, yielding a score in $[-1, +1]$ analogous to VADER's compound score, where negative values indicate negative sentiment and positive values indicate positive sentiment. The BERT Weighted Average was computed as the dot product with $[1, 2, 3, 4, 5]$ divided by 5, producing a normalized score in $[0.2, 1.0]$. Tracks with fewer than 50 words in the original lyrics were excluded to ensure sufficient textual input for reliable transformer inference, and input texts were truncated at the model's maximum context window of 512 tokens; for longer songs, only the opening portion was scored. To quantify the impact of this truncation, a robustness check was performed in which every track was additionally scored by splitting its full lyrics into $\leq$500-token chunks, scoring each chunk independently, and averaging the per-chunk composite scores; results are reported in Section 4.3.1. All inference was performed in evaluation mode (no gradient

computation) using PyTorch 2.6.0+cu124 and Hugging Face Transformers 4.57.3, with automatic device selection.

A separate preprocessing step generated a cleaned lyrics column for use in the subsequent fractal and complexity analyses (Section 3.7). This cleaning involved lowercasing, removal of URLs and non-alphabetic characters (preserving language-specific accented characters), elimination of repeated word patterns, and removal of multilingual stopwords (English, Spanish, and German) including domain-specific vocalizations common in song lyrics (e.g., "yeah," "oh," "uh"). The cleaned text was retained in the output dataset but was not used as input to the BERT model.

External Validation on Lyric-Domain Mood Annotations

Because the multilingual BERT model used here was fine-tuned on product reviews rather than song lyrics, an additional external validation was performed against an independent, lyric-specific gold standard. We used the MoodyLyrics dataset [31], which provides mood annotations for 2595 English-language tracks distributed across four categories (happy, relaxed, sad, angry) drawn from the four quadrants of Russell's valence–arousal model. Lyrics were retrieved via the Genius API; 2485 tracks (95.8%) returned a match, of which 2444 met the same $\geq$50-word threshold used for the main analysis. Each track was scored with the same BERT pipeline (Section 3.6), and the BERT composite score was compared against the MoodyLyrics valence label, defined as positive for happy and relaxed and as negative for sad and angry. Two complementary analyses were conducted. First, we computed binary classification metrics treating the BERT composite as a positive/negative classifier (composite > 0 predicts positive). Second, we treated the BERT composite as a continuous score and computed its rank correlation (Spearman ρ) with the four-class mood label and the point-biserial correlation with the binary valence label. Per-quadrant means provided a fine-grained check on whether BERT's score ordering matched the underlying valence–arousal structure of the gold standard.

3.7. Zipf's Law and Power-Law Analysis

The word frequency distributions of the lyrics corpus were analyzed for adherence to Zipf's law, which predicts a power-law relationship between word rank and frequency such that the frequency of the r-th ranked word is proportional to $r^{-\alpha}$ [12]. For each language subcorpus, lyrics were tokenized by lowercasing, removing non-alphabetic characters (preserving language-specific accented characters: ä, ö, ü, ß, á, é, í, ó, ú, ñ), and filtering tokens shorter than two characters. Word frequencies were computed and sorted in descending order, and the Zipf exponent α was estimated via ordinary least squares (OLS) regression of $\log_{10}(\text{frequency})$ on $\log_{10}(\text{rank})$. Goodness of fit was assessed by R^2 and the standard error of the slope estimate.

To supplement the OLS Zipf analysis, maximum likelihood estimates of the scaling exponent α and lower bound $x_{\min}$ were obtained using the `powerlaw` Python package [32]. The MLE estimates are reported alongside the OLS estimates, but likelihood-ratio comparisons against alternative distributions were not retained because the corresponding diagnostic outputs were unavailable. Additionally, vocabulary richness was quantified using four complementary metrics: the type–token ratio (TTR), Herdan's C (a log-corrected variant of TTR that partially controls for corpus size), Brunet's W, and the hapax legomena ratio (proportion of word types occurring exactly once). These metrics provide converging evidence on lexical diversity independent of the Zipf exponent and are less sensitive to differences in subcorpus size than the raw TTR.

3.8. Multifractal Detrended Fluctuation Analysis and Long-Range Dependence

To characterize the scaling and long-range dependence properties of the lyrics corpus, detrended fluctuation analysis (DFA; [16]) and its multifractal generalization (MF-DFA; [17]) were applied to word-level time series constructed from each language subcorpus. Two types of series were analyzed: a word-length series, in which each word was represented by its character count, and a word frequency rank series, in which each word was replaced by the natural logarithm of its corpus-wide frequency rank. These representations capture complementary structural aspects of text: the word-length series reflects phonotactic and morphological patterns, while the rank series captures the sequential deployment of common versus rare vocabulary.

Standard DFA estimates a single Hurst exponent H from the scaling of the root-mean-square fluctuation function $F(s)$ across segment sizes s, where $H > 0.5$ indicates persistent long-range correlations and $H = 0.5$ corresponds to an uncorrelated random process. The cumulative profile $Y(i)$ was computed by subtracting the series mean and applying a cumulative sum. Segments were detrended using a first-order polynomial fit, and the fluctuation function was computed across 20 logarithmically spaced scales. H was estimated as the slope of $\log_2(F)$ versus $\log_2(s)$ via ordinary least squares.

MF-DFA extends this framework by computing a generalized fluctuation function $F_q(s)$ for moment orders q from -5 to $+5$ in steps of 0.5, yielding the generalized Hurst exponent $h(q)$. The singularity spectrum $f(\alpha)$ was obtained via a Legendre transform, and the spectrum width $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$ was used as the primary measure of multifractality. Two validation procedures were employed: (1) a surrogate shuffle test ($N = 50$ permutations per series) to confirm that $H > 0.5$ reflects genuine long-range dependence rather than distributional artifacts, and (2) a bootstrap subsampling test (100 iterations, $n = 225$ songs per language) to control for unequal subcorpus sizes when comparing $\Delta\alpha$ across languages, with pairwise Mann–Whitney U tests for statistical comparison.

3.9. Streaming Distribution Scaling and Box-Counting Fractal Dimension

The streaming counts available in the present corpus reflect a single weekly snapshot per track (the first occurrence retained during deduplication; Section 3.1) drawn from the Top 200 chart only. They therefore do not represent peak, cumulative, or average popularity, and they truncate the long tail of less-streamed releases that never entered the Top 200. The analyses in this section should accordingly be understood as descriptors of the internal concentration of weekly streaming activity within the Top 200, not as estimates of the population-level distribution of song popularity. Within this restricted scope, we regressed $\log_{10}(\text{streams})$ on $\log_{10}(\text{rank})$ within each language subcorpus to obtain a within-chart streaming exponent for descriptive comparison with the lexical Zipf exponent estimated in Section 3.7. Concentration of streaming activity was further quantified using the Gini coefficient, where values closer to 1 indicate greater inequality (i.e., a small number of tracks capturing a disproportionate share of streams). Additionally, box-counting fractal dimension was computed on two representations: (1) the one-dimensional distribution of word lengths (character counts per token), and (2) the two-dimensional log-rank versus log-frequency scatter of word types. For each representation, the number of occupied boxes $N(\varepsilon)$ was counted across 20 logarithmically spaced box sizes ε, and the fractal dimension D was estimated as the slope of $\log(N)$ versus $\log(1/\varepsilon)$ via ordinary least squares.

4. Results

4.1. Dataset Characteristics

The deduplicated full corpus comprised 2113 unique tracks by 672 artists, extracted from 10,800 chart entries across 54 weekly Spotify Top 200 snapshots spanning seven regional markets. After language filtering (Section 3.3), the analytical corpus restricted to English, Spanish, and German contained 2023 tracks (95.74% of the deduplicated full corpus). The temporal distribution of snapshots was skewed toward earlier years (2019: $n = 31$; 2020: $n = 17$; 2021: $n = 6$), reflecting limited Wayback Machine archival availability for spotifycharts.com beyond early 2021. Table 2 reports the linguistic composition and lyrics coverage of the corpus. Table 3 summarizes the weekly stream-count distribution for the deduplicated corpus.

Table 2. Linguistic composition and lyrics coverage of the deduplicated full corpus.

Language	*n*	% of Corpus	LRCLIB	Genius	VADER Matched
English	1491	70.56	—	—	1312 (88.0%)
Spanish	307	14.53	—	—	246 (80.1%)
German	225	10.65	—	—	213 (94.7%)
Other/unknown	90	4.26	—	—	—
Total	2113	100.00	2098 (99.29%)	1904 (90.11%)	1833 (86.7%)

Note. The "Other/unknown" row includes tracks assigned to languages outside English, Spanish, and German, as well as tracks classified as unknown because lyrics were not retrieved. LRCLIB and Genius columns report tracks with lyrics retrieved. VADER Matched reports tracks successfully linked across data sources after encoding normalization (1798 via accent-stripped matching + 35 via aggressive normalization). Per-language VADER match rates are for the top-3 language subset ($N = 2023$).

Table 3. Descriptive statistics of weekly stream counts (deduplicated corpus; $N = 2113$).

Min	Max	*M*	*SD*	*Mdn*	*N*
284,556	65,873,080	6,869,596	5,937,902	5,567,732	2113

Note. Stream counts correspond to a single weekly chart snapshot per track (first occurrence retained during deduplication) and do not represent cumulative or peak values.

4.2. Baseline Sentiment Characteristics (VADER)

VADER sentiment analysis was performed on the 1904 tracks for which Genius lyrics were available. Of the 1833 Genius-scored tracks successfully linked to the LRCLIB-based corpus, classification yielded 922 positive tracks (50.3%) and 911 negative tracks (49.7%) across all detected languages, indicating a balanced overall sentiment distribution. Table 4 reports the VADER compound sentiment scores stratified by language for the top-3 language subset ($n = 1771$).

Table 4. VADER compound sentiment scores by language (top-3 language subset).

Language	*n*	*M*	*SD*	*Mdn*	Min	Max
English	1312	0.2601	0.8822	0.9307	−1.0000	0.9999
Spanish	246	−0.5798	0.6673	−0.9315	−0.9975	0.9993
German	213	−0.7426	0.5140	−0.9757	−0.9994	0.9986

Note. Compound scores computed using VADER [8] on Genius-sourced lyrics. VADER is an English-language lexicon; cross-linguistic differences should be interpreted with caution (see Section 4.3).

As shown in Figure 2 and Table 4, substantial cross-linguistic differences were observed in VADER compound sentiment. English-language lyrics exhibited a bimodal distribution with a strong positive peak near +1.0 ($M = 0.260$, $Mdn = 0.931$), whereas Spanish

($M = -0.580$, $Mdn = -0.932$) and German ($M = -0.743$, $Mdn = -0.976$) lyrics were concentrated in the negative range. However, because VADER is an English-centric lexicon that does not recognize sentiment-bearing vocabulary in other languages, the strongly negative scores for Spanish and German likely reflect a measurement artifact rather than genuinely more negative lyrical content. This limitation motivates the use of a multilingual transformer-based sentiment model (Section 3.6).

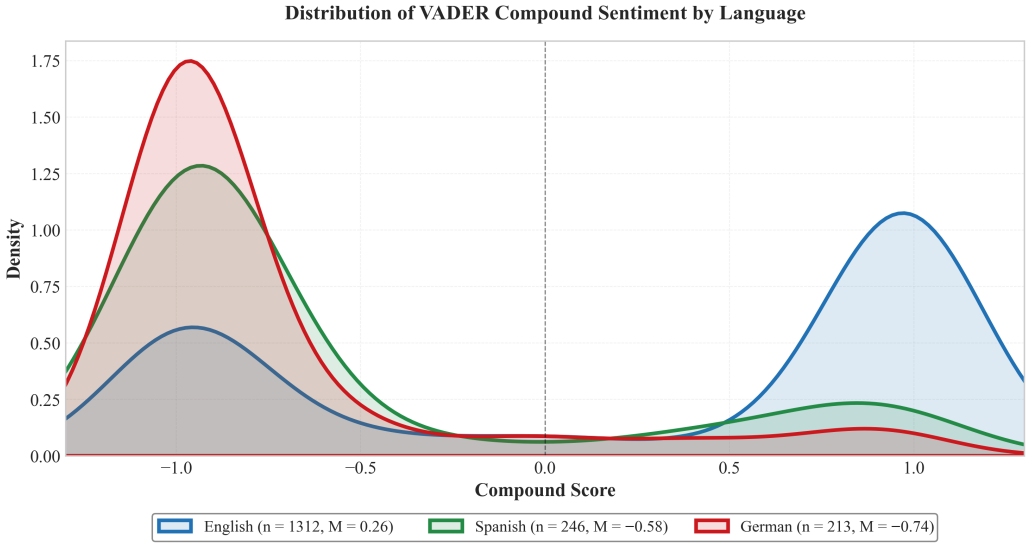


Figure 2. Kernel density estimates of VADER compound sentiment scores by language (English, Spanish, German). The dashed vertical line indicates the neutral threshold (0.0). Sample sizes and means are reported in the legend.

4.3. BERT Sentiment Results

BERT sentiment scores were computed for 2019 of the 2023 tracks in the top-3 language subset (4 tracks excluded due to insufficient lyrics length). Across all languages, the overall mean BERT Composite Score was negative, with 1287 tracks (63.7%) receiving a negative composite score and 732 (36.3%) receiving a positive score. The mean star-class probability distribution was dominated by the extreme classes: 1-star (31.1%) and 5-star (23.2%), with intermediate classes receiving lower probabilities (2-star: 19.2%; 3-star: 12.9%; 4-star: 13.6%), indicating a polarized sentiment landscape consistent with the affective intensity typical of popular song lyrics. Table 5 reports the BERT Composite Score by language.

Table 5. BERT Composite Sentiment Scores by language (top-3 language subset).

Language	*n*	*M*	*SD*	*Mdn*	Min	Max	% Negative
English	1489	−0.0920	0.3719	−0.1134	−0.8883	0.9521	61.3%
Spanish	305	−0.2034	0.3274	−0.2419	−0.9071	0.7141	73.8%
German	225	−0.0760	0.3034	−0.0995	−0.7588	0.8942	66.2%

Note. Composite scores computed using `nlptown/bert-base-multilingual-uncased-sentiment` on original (non-preprocessed) lyrics. % Negative indicates the proportion of tracks with a composite score below 0. The BERT model was fine-tuned on multilingual product reviews; cross-linguistic comparisons are therefore more reliable than the VADER baseline (Table 4).

4.3.1. Robustness to 512-Token Truncation

The multilingual BERT model has a maximum input length of 512 tokens, and the BERT scores reported above were computed on the opening 512-token window of each track. Because 1329 of the 2019 scored tracks (65.8%) exceeded this length—disproportionately so for Spanish (255/305, 83.6%) and German (186/225, 82.7%) lyrics, which were on average

longer than English—the opening-only scores omit material from the second half of most songs. To assess whether this truncation affected the cross-linguistic conclusions, we re-scored every track by tokenizing the full lyrics, splitting them into non-overlapping chunks of $\leq$500 tokens, scoring each chunk with the same BERT model, and averaging the per-chunk composite scores to obtain a chunk-averaged "full-song" composite (median chunks per track = 2 in all three languages). Table 6 compares the opening-window and chunk-averaged scores.

Table 6. Robustness check: opening-window vs. chunk-averaged BERT composite scores.

Lang.	*n*	Opening *M* (*SD*)	Chunked *M* (*SD*)	Pearson *r*	Mean $\lvert\Delta\rvert$	Sign Agree.
English	1489	−0.092 (0.372)	−0.100 (0.361)	0.909	0.097	89.1%
Spanish	305	−0.203 (0.328)	−0.206 (0.301)	0.874	0.113	88.2%
German	225	−0.076 (0.304)	−0.081 (0.305)	0.864	0.117	82.7%
Overall	2019	—	—	0.902	0.102	88.3%

Note. Opening-window scores are the original BERT composite scores reported in Table 5 (computed on the first 512 tokens of each track). Chunk-averaged scores are the mean composite across all $\leq$500-token chunks of the full lyrics. Pearson *r*, mean absolute difference, and sign agreement are computed track-wise between the two scores. "Sign agreement" is the proportion of tracks for which both scores share the same sign (i.e., agree on positive vs. negative valence).

The two scoring strategies were highly concordant. Track-wise Pearson correlations between the opening-window and chunk-averaged composites were $r = 0.909$ for English, $r = 0.874$ for Spanish, and $r = 0.864$ for German (all $p < 10^{-50}$); the mean absolute difference between the two scores was 0.10 on a $[-1, +1]$ scale; and the two strategies agreed on positive/negative valence for 88.3% of tracks overall (English 89.1%; Spanish 88.2%; German 82.7%). Of the 237 tracks (11.7%) whose valence sign flipped between the two scoring strategies, the imbalance was slightly higher for German (17.3% flips), consistent with German's higher proportion of long, multi-chunk tracks. Critically, the by-language mean composite scores changed by less than 0.01 in all three languages (English $-0.092 \rightarrow -0.100$; Spanish $-0.203 \rightarrow -0.206$; German $-0.076 \rightarrow -0.081$), preserving the central finding that the cross-linguistic gap is small ($\leq$0.13) regardless of scoring strategy. We therefore retain the opening-window scores as the primary BERT measurement throughout the manuscript, while noting that all cross-linguistic conclusions are stable under the more conservative full-song chunk-averaged alternative. The track-wise relationship is tight and approximately diagonal: scores cluster around the identity line $y = x$ with no systematic shift, the bulk of points falls in the lower-left negative-valence quadrant in all three languages, and disagreements are concentrated near the neutral threshold (composite $\approx$ 0) where small numerical differences flip the sign without altering the substantive interpretation.

The BERT results substantially revised the cross-linguistic sentiment pattern observed with VADER. Under VADER, Spanish and German lyrics appeared overwhelmingly negative ($M = -0.580$ and -0.743, respectively), while English was predominantly positive ($M = 0.260$). The multilingual BERT model revealed a markedly different picture: all three languages exhibited mildly negative mean composite scores, with German ($M = -0.076$, $SD = 0.304$) and English ($M = -0.092$, $SD = 0.372$) nearly indistinguishable, and Spanish slightly more negative ($M = -0.203$, $SD = 0.328$). The cross-linguistic gap narrowed from a 1.003-point spread under VADER (English *M* minus German *M*) to a 0.127-point spread under BERT, confirming that the extreme VADER differences were largely attributable to the lexicon's inability to recognize sentiment-bearing vocabulary in non-English languages rather than genuine affective differences in lyrical content.

At the classification level, BERT identified 61.3% of English tracks as negatively valenced (composite < 0), compared to 36.6% under VADER. For Spanish, BERT classified 73.8% as negative versus 84.0% under VADER; for German, 66.2% versus 88.3%. Thus, while BERT also detected a slight negative skew in popular music lyrics across all three languages, the magnitude of cross-linguistic differences was dramatically reduced. Figure 3 presents kernel density estimates of the BERT Composite Score distributions by language.

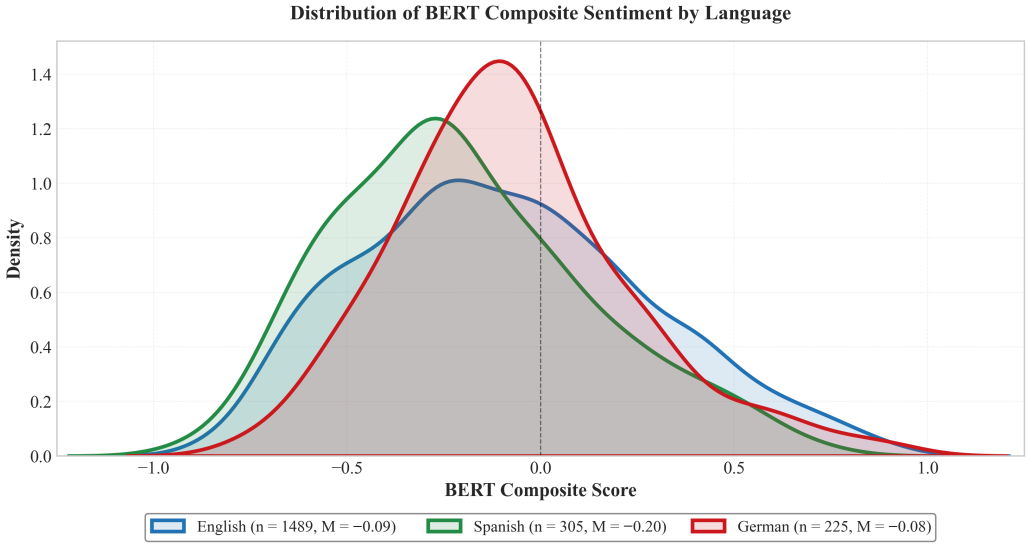


Figure 3. Kernel density estimates of BERT Composite Sentiment scores by language (English, Spanish, German). The dashed vertical line indicates the neutral threshold (0.0). Sample sizes and means are reported in the legend.

4.3.2. External Validation of BERT on Lyric-Domain Mood Annotations

To assess whether the multilingual BERT model used here generalizes from its product-review training domain to the affective vocabulary of song lyrics, we evaluated it against the MoodyLyrics gold standard [31]. Of the 2595 MoodyLyrics tracks, 2444 were successfully retrieved via Genius and met the $\geq$50-word threshold (94.2%); the remaining tracks were either not found on Genius or returned lyrics shorter than 50 words. Treating the BERT composite as a binary positive/negative classifier against the MoodyLyrics valence label, BERT achieved an accuracy of 71.5%, with precision 0.71, recall 0.81, and F_1 score 0.76 for the positive class (Table 7). The receiver operating characteristic area under the curve was AUC $= 0.795$, indicating substantially above-chance discrimination.

Stratifying by MoodyLyrics quadrant (Table 8) shows that the BERT composite score ordering matched the expected valence ranking: happy ($M = 0.456$) $>$ relaxed ($M = 0.263$) $>$ angry ($M = -0.066$) $>$ sad ($M = -0.166$). This monotonic ordering across all four quadrants—combined with the strong rank correlation ($\rho = 0.509$, $p < 10^{-150}$)—indicates that BERT's continuous score reflects a graded valence dimension consistent with how human annotators labeled these lyrics, even though the model was not trained on song-lyric data.

These accuracy and correlation values fall within the range typically reported for general-purpose sentiment models applied to lyric-domain benchmarks [3,31]. While a domain-adapted classifier fine-tuned on song lyrics would likely outperform the present model, the validation confirms that the cross-linguistic comparisons reported in Section 4.3 are based on a sentiment signal that is genuinely informative for popular song lyrics and not an artifact of out-of-domain inference.

Table 7. External validation of BERT on the MoodyLyrics gold standard ($n = 2444$).

Metric	Value
Accuracy	0.715
Precision (positive)	0.71
Recall (positive)	0.81
F_1 (positive)	0.76
ROC AUC	0.795
Point-biserial r	0.506 ($p < 10^{-150}$)
Spearman ρ (vs. 4-class mood)	0.509 ($p < 10^{-150}$)
Confusion matrix	
True negatives	670
False positives	440
False negatives	257
True positives	1077

Note. Positive class = happy or relaxed; negative class = sad or angry. BERT prediction is positive if the composite score > 0.

Table 8. BERT composite score by MoodyLyrics quadrant.

Mood	*n*	BERT *M*	BERT *SD*	BERT *Mdn*	Star *M*
happy	772	0.456	0.393	0.544	4.46
relaxed	562	0.263	0.404	0.304	3.96
angry	548	−0.066	0.420	−0.062	2.84
sad	562	−0.166	0.429	−0.189	2.47

Note. Star *M* = mean predicted star rating (1–5) from the underlying BERT classifier.

Figure 4 visualizes the corresponding binary confusion matrix.

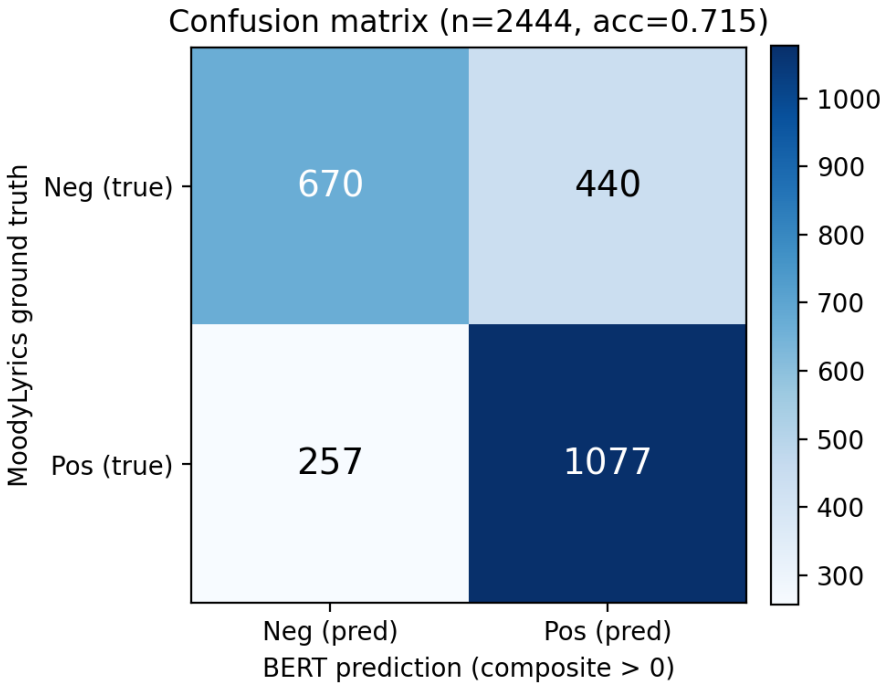


Figure 4. Confusion matrix for BERT binary valence prediction (composite score > 0) against MoodyLyrics ground-truth labels (positive = happy or relaxed; negative = sad or angry). Accuracy = 71.5%; $n = 2444$.

4.4. Zipf's Law and Lexical Diversity

Word frequency distributions for all three language subcorpora exhibited strong adherence to Zipf's law, with R^2 values exceeding 0.96 across all languages. Table 9

reports the Zipf exponents estimated via OLS log–log regression alongside corpus statistics. English lyrics yielded the steepest exponent ($\alpha = 1.409$, $SE = 0.001$, $R^2 = 0.982$), followed by Spanish ($\alpha = 1.220$, $SE = 0.002$, $R^2 = 0.979$) and German ($\alpha = 1.181$, $SE = 0.002$, $R^2 = 0.969$). All exponents exceeded the classical Zipfian value of $\alpha = 1.0$, consistent with the constrained vocabulary typically observed in popular song lyrics relative to natural prose. The higher English exponent indicates greater lexical concentration, with a small set of high-frequency words dominating the corpus more strongly than in the other two languages.

Table 9. Zipf exponents and corpus statistics by language.

Language	*n* Songs	Tokens	Types	α_{MLE}	α_{OLS}	R^2	*SE*
English	1491	594,703	19,961	1.556	1.409	0.982	0.001
Spanish	307	133,643	9908	1.592	1.220	0.979	0.002
German	225	92,029	9083	1.633	1.181	0.969	0.002

Note. α_{MLE} = maximum likelihood estimate via the Clauset–Shalizi–Newman method ($x_{\min} = 1$ for all languages). α_{OLS} = ordinary least squares estimate from log–log regression. R^2 and *SE* refer to the OLS fit. Tokens and types are based on lowercased, minimally filtered text (tokens ≥ 2 characters).

Vocabulary richness metrics corroborated the cross-linguistic pattern suggested by the Zipf exponents (Table 10). German lyrics exhibited the highest type–token ratio (TTR = 0.099), Herdan's C (0.797), and hapax legomena ratio (50.3%), indicating the greatest lexical diversity. Spanish occupied an intermediate position (TTR = 0.074, Herdan's C = 0.780, hapax ratio = 42.8%), while English showed the lowest diversity (TTR = 0.034, Herdan's C = 0.745, hapax ratio = 43.3%). The elevated German diversity is consistent with the language's productive compound word formation system, which generates a larger number of unique word forms. The inverse relationship between Zipf exponent and vocabulary richness—English had the steepest exponent but lowest TTR—confirms that English-language lyrics are characterized by heavier reliance on a small set of high-frequency words.

Table 10. Vocabulary richness metrics by language.

Language	TTR	Herdan's C	Brunet's W	Hapax Ratio	Hapax (n)
English	0.034	0.745	11.263	43.3%	8651
Spanish	0.074	0.780	11.300	42.8%	4244
German	0.099	0.797	10.843	50.3%	4571

Note. TTR = type–token ratio. Herdan's C = $\log(\text{types})/\log(\text{tokens})$. Brunet's W = $\text{tokens}^{(\text{types}^{-0.172})}$. Hapax ratio = proportion of word types occurring exactly once. Raw TTR is sensitive to corpus size; Herdan's C provides a partially size-corrected alternative.

Figure 5 presents the log–log rank–frequency plots for each language, with OLS regression lines overlaid. The linearity of the fits across approximately four orders of magnitude confirms that the word frequency distributions are well-described by a power-law model. The systematic deviation from linearity in the low-rank (high-frequency) tail is a well-known departure from strict Zipf behavior, often attributed to the disproportionate frequency of function words [13].

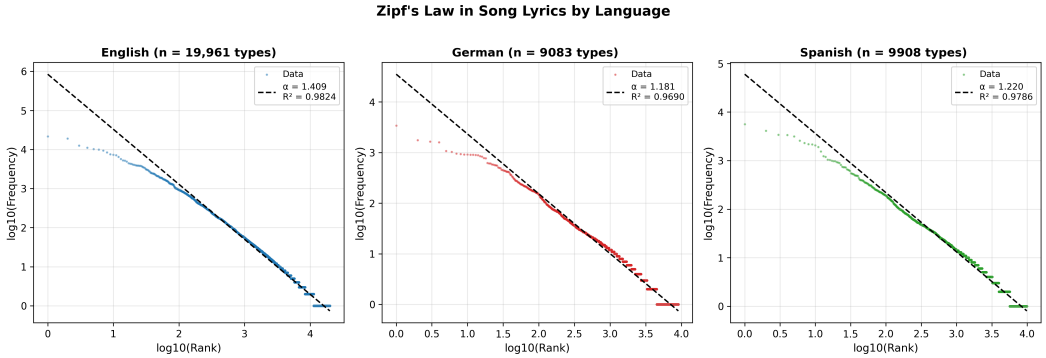


Figure 5. Zipf's law log–log rank–frequency plots for English, German, and Spanish lyrics. Black dashed lines indicate OLS regression fits. Zipf exponents (α) and R^2 values are reported in the legends.

4.5. Multifractal Structure and Long-Range Dependence

DFA revealed persistent long-range correlations in all language subcorpora and both series types (Table 11). All Hurst exponents substantially exceeded the $H = 0.5$ threshold, ranging from $H = 0.664$ (Spanish word-length) to $H = 0.762$ (English frequency-rank). The word frequency rank series consistently yielded higher H values than the word-length series, suggesting that the sequential ordering of common versus rare words exhibits stronger long-range memory than the pattern of word lengths. All DFA fits achieved $R^2 > 0.995$.

Table 11. DFA Hurst exponents by language and series type.

Language	H (Word-Length)	R^2	H (Freq-Rank)	R^2
English	0.687	0.997	0.762	0.997
Spanish	0.664	0.998	0.671	0.996
German	0.691	0.996	0.736	0.996

Note. H = Hurst exponent from DFA with first-order polynomial detrending. $H = 0.5$ indicates uncorrelated; $H > 0.5$ indicates persistent long-range correlations. The freq-rank H values equal $h(q = 2)$ from MF-DFA (Table 12).

The surrogate shuffle test confirmed that these elevated Hurst exponents reflect genuine temporal structure. Across all six language $\times$ series-type combinations, shuffled surrogates produced H values tightly centered on 0.50 (range of means: 0.498–0.502, $SD \leq 0.013$), while observed H values (0.664–0.762) fell entirely outside the surrogate distributions; no shuffled surrogate exceeded the corresponding observed value in any case (0/50 surrogates). Figure 6 presents the surrogate distributions alongside observed values.

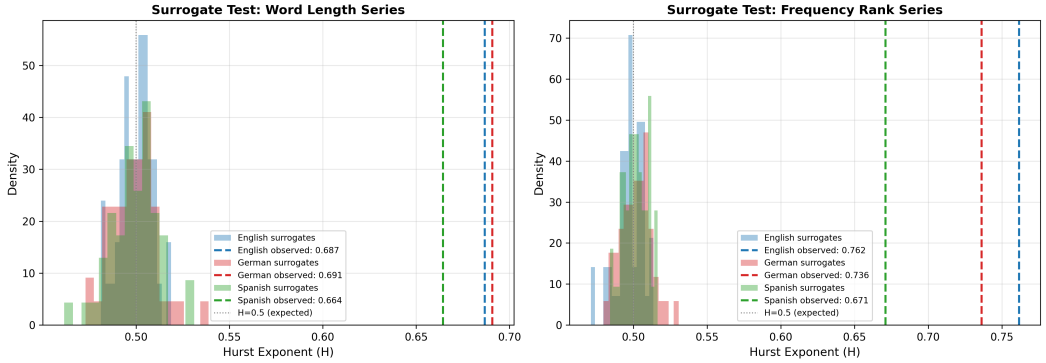


Figure 6. Surrogate shuffle test for DFA Hurst exponents. (**Left**): word-length series. (**Right**): frequency rank series. Histograms show H from 50 shuffled surrogates; dashed lines indicate observed H. Dotted gray line marks $H = 0.5$.

MF-DFA applied to the full word frequency rank series revealed substantial multifractal structure in all three languages (Table 12). The singularity spectrum width $\Delta\alpha$ was largest for Spanish (3.321), followed by German (2.985) and English (2.466). The $h(q = 2)$ values matched the DFA frequency-rank Hurst exponents, confirming internal consistency. Figure 7 presents the singularity spectra.

Table 12. MF-DFA multifractal spectrum parameters by language.

Language	$\Delta\alpha$ (Corpus)	$h(q = 2)$	$\Delta\alpha$ Boot. M	$\Delta\alpha$ Boot. SD	$\Delta\alpha$ Boot. Mdn
English	2.466	0.762	2.591	0.932	2.648
Spanish	3.321	0.671	2.736	0.891	2.810
German	2.985	0.736	2.740	0.108	2.735

Note. $\Delta\alpha$ (corpus) = singularity spectrum width on full subcorpus. $h(q{=}2)$ = generalized Hurst exponent at $q = 2$, equivalent to DFA H (Table 11, freq-rank). Bootstrap: 100 resamples of $n = 225$ songs per language.

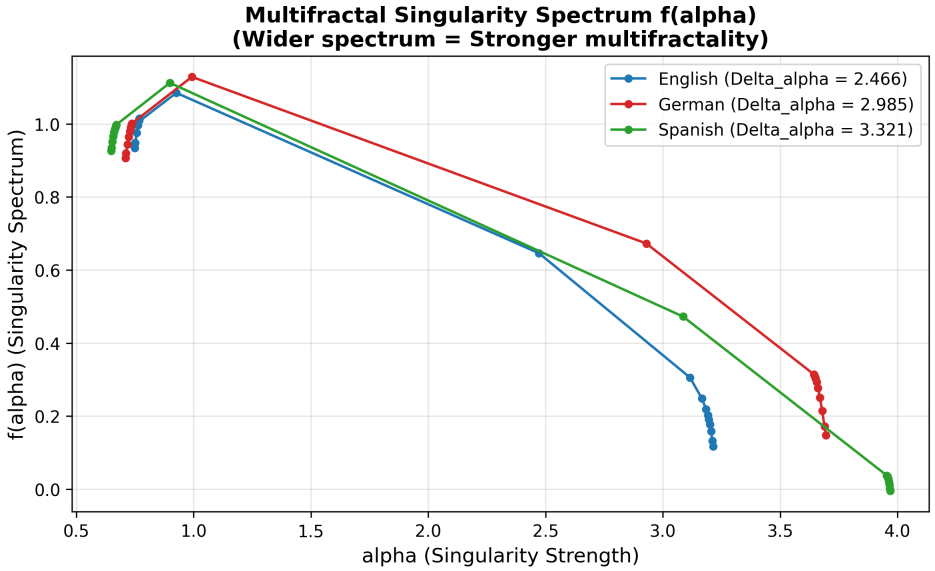


Figure 7. Multifractal singularity spectra $f(\alpha)$ for the word frequency rank series by language. Wider spectra indicate stronger multifractality. Spectrum widths ($\Delta\alpha$) are reported in the legend.

However, the bootstrap subsampling test revealed that these apparent cross-linguistic differences were largely attributable to unequal subcorpus sizes. When each language was subsampled to $n = 225$ songs (100 iterations), the $\Delta\alpha$ distributions converged: English $M = 2.591$ ($SD = 0.932$, $Mdn = 2.648$), Spanish $M = 2.736$ ($SD = 0.891$, $Mdn = 2.810$), and German $M = 2.740$ ($SD = 0.108$, $Mdn = 2.735$). Pairwise Mann–Whitney U tests found no significant differences (English vs. German: $U = 4391$, $p = 0.137$; English vs. Spanish: $U = 4663$, $p = 0.411$; German vs. Spanish: $U = 4462$, $p = 0.189$). Figure 8 presents the bootstrap distributions. The narrow German variance ($SD = 0.108$) reflects minimal sampling variability, as 225 songs constitutes the full German subcorpus.

Taken together, these results indicate that popular song lyrics across the three languages examined here exhibit multifractal scaling and persistent long-range dependence, with Hurst exponents $H \approx 0.66$–0.76 and spectrum widths $\Delta\alpha \approx 2.6$–2.7 when corpus size is controlled. Within this restricted three-language set—all Indo-European—multifractal properties appear to be a shared rather than language-specific feature, consistent with (but not yet establishing) the interpretation that songwriting constraints such as melodic structure, rhyme schemes, and verse–chorus repetition impose similar long-range organizational patterns. Whether this pattern generalizes to typologically distant languages (e.g., tonal, agglutinative, or non-Indo-European) cannot be determined from the present data and remains an open empirical question.

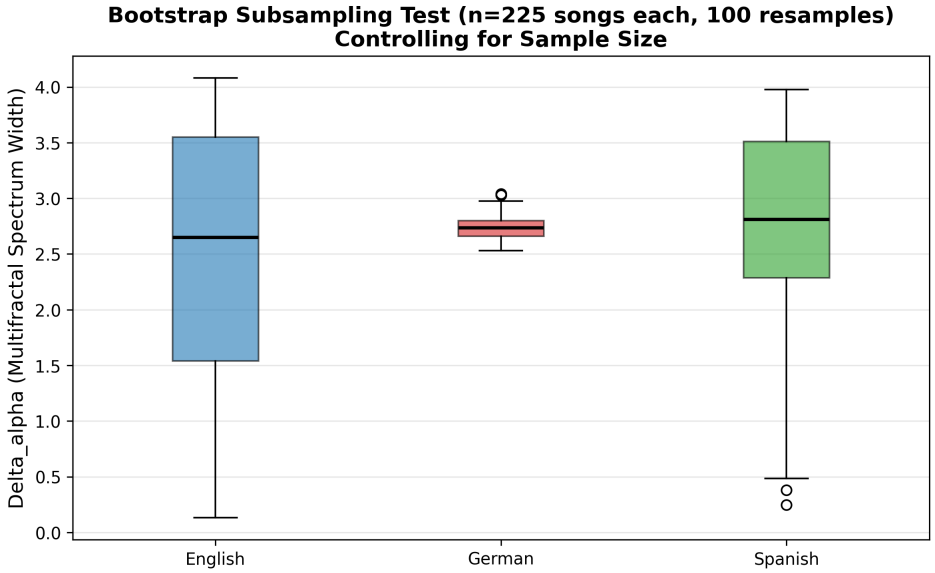


Figure 8. Bootstrap subsampling test for $\Delta\alpha$. Each language was subsampled to $n = 225$ songs (100 iterations). No pairwise differences were significant (all $p > 0.13$).

4.6. Within-Chart Streaming Concentration and Fractal Dimension

As detailed in Section 3.9, the streaming counts analyzed here reflect a single weekly Top 200 snapshot per track and capture only the head of the popularity distribution; the results below should therefore be read as descriptors of within-chart streaming concentration rather than as evidence of a population-level power law. The within-chart rank–frequency relationship was approximately log-linear in all three languages, but with substantially lower R^2 values than the lexical Zipf fits, consistent with the truncated and snapshot-based nature of the sample. German showed the steepest within-chart exponent ($\alpha = 0.994$, $SE = 0.024$, $R^2 = 0.886$) and the highest Gini coefficient (0.556), indicating the most concentrated within-chart listening pattern, in which a small number of top-charting tracks accounted for a disproportionate share of weekly streams. English exhibited an intermediate exponent ($\alpha = 0.744$, $SE = 0.016$, $R^2 = 0.602$, Gini $= 0.390$), while Spanish had the shallowest slope ($\alpha = 0.542$, $SE = 0.024$, $R^2 = 0.627$, Gini $= 0.298$), consistent with a more evenly distributed within-chart consumption pattern. Given the sampling design, these exponents and Gini values should not be interpreted as estimates of population-level popularity scaling and are not used to support claims about heavy-tailed popularity distributions of the kind reported for unbounded online diffusion data [24]. Box-counting fractal dimension of the word-length distribution yielded $D \approx 0.42$–0.45 across languages ($R^2 \approx 0.78$–0.80), reflecting the discrete, bounded nature of word lengths (integers typically ranging from 2 to 15 characters). The two-dimensional rank–frequency fractal dimension was $D \approx 0.92$–0.94 ($R^2 > 0.997$) for all languages, confirming that the log-transformed Zipf scatter approximates a one-dimensional curve in two-dimensional space—a result that is geometrically expected given the near-linear log–log relationship established by Zipf's law and does not provide independent information beyond the Zipf exponents already reported.

5. Discussion

This study set out to bridge three largely independent research traditions—sentiment analysis, statistical physics of language, and music information retrieval—in a single cross-linguistic analysis of popular song lyrics. The results yield several findings with implications for each of these fields.

The most striking result concerns the dramatic revision of cross-linguistic sentiment patterns when moving from the English-centric VADER lexicon to a multilingual BERT model. The 1.003-point gap between English and German under VADER collapsed to just 0.127 points under BERT, confirming that the extreme negativity scores previously attributed to non-English lyrics were largely measurement artifacts arising from VADER's inability to recognize sentiment-bearing vocabulary outside English. This finding carries a cautionary methodological message for the broader computational musicology literature: studies that have applied English-centric lexicons to multilingual corpora without cross-validating against language-equitable models may have reported spurious cross-linguistic differences [3,25]. The residual cross-linguistic variation under BERT—with Spanish lyrics slightly more negative than English or German—may reflect genuine cultural or genre-level differences in lyrical content, though domain transfer limitations of the BERT model (fine-tuned on product reviews rather than song lyrics) prevent strong conclusions on this point [10].

The complexity-science results reveal a striking degree of cross-linguistic consistency in the structural organization of popular music lyrics within the three languages examined here. English, Spanish, and German all exhibited Zipfian word frequency distributions ($R^2 > 0.96$), persistent long-range correlations ($H \approx 0.66$–0.76; none of the 50 shuffled surrogates exceeded the observed values), and broad multifractal spectra ($\Delta\alpha \approx 2.5$–3.3). Critically, the bootstrap subsampling test demonstrated that the apparent cross-linguistic differences in multifractal spectrum width were not statistically significant when corpus size was controlled (all pairwise $p > 0.13$), suggesting that, within this set, multifractal scaling is a shared rather than language-specific structural property of commercially successful lyrics. This extends the findings of Montemurro and Pury [18] and Drożdż et al. [19], who documented similar properties in literary prose, to the more constrained register of song lyrics, and is consistent with the interpretation that songwriting constraints—melodic structure, verse–chorus repetition, rhyme schemes—impose long-range organizational patterns that transcend individual languages. We emphasize, however, that the three languages analyzed here all belong to the Indo-European family and that English is heavily over-represented relative to Spanish and German; whether the same patterns hold in typologically distant languages, in non-chart popular music, or in larger and more balanced corpora cannot be determined from the present data.

Where the languages did differ was in lexical diversity: German exhibited the shallowest Zipf exponent ($\alpha = 1.181$), highest type–token ratio (0.099), and largest hapax ratio (50.3%), while English showed the steepest exponent ($\alpha = 1.409$) and lowest diversity. These differences are consistent with German's productive compound word formation system [13] and suggest that cross-linguistic variation in lyrics is better explained by linguistic typology than by the constraints of songwriting per se.

Read together, the structural and lexical findings invite a more concrete interpretation of what the chart-lyric register actually looks like as a written form. The Zipf exponents observed here ($\alpha = 1.18$–1.41) sit clearly above the value of approximately 1.0 typically reported for literary prose [13,21], meaning that a small number of high-frequency words carries proportionally more of the text than in long-form writing. Operationally, this is the statistical signature of the chorus-driven structure of popular songs: a short refrain that recurs many times, intersected with shorter verses that introduce new vocabulary at a much lower rate. The same compositional fact appears, from a different angle, in the persistent long-range correlations measured by DFA ($H \approx 0.66$–0.76). A Hurst exponent in this range means that the choice of word at one point in a song is statistically informative about word choices several stanzas away—i.e., the lyrics are not locally independent but cohere as an extended sequence, even after controlling for surface repetition. The broad

multifractal spectra ($\Delta\alpha \approx 2.5$–3.3) indicate that this temporal coherence is not produced by a single mechanism: rare and frequent words follow distinct scaling regimes, consistent with songs in which a stable lexical core (function words, refrain vocabulary) coexists with thematically clustered open-class words that arrive in bursts. The cross-linguistic divergence in lexical diversity admits a similarly concrete reading: German's higher hapax ratio (50.3%, against 43.3% for English) is unlikely to reflect more "varied" songwriting in any aesthetic sense, and is more parsimoniously explained by morphology—German's productive nominal compounding inflates the type count for what would, in English, be expressed as multiword phrases. From the standpoint of language use, the joint pattern of steep Zipf slopes, persistent long-range memory, and broad multifractality places chart lyrics close to literary prose in their structural organization while remaining lexically far more compressed, supporting the view that the lyric register is not a degenerate or impoverished form of language but a constrained register whose departures from prose are interpretable in terms of compositional function rather than linguistic deficit.

Several limitations should be acknowledged. First, the corpus, while substantial (2023 tracks), is drawn exclusively from Spotify Top 200 charts across a limited temporal window (effectively 2019–2021), and may not generalize to non-chart music, historical periods, or languages beyond English, Spanish, and German. Relatedly, the streaming counts retained for each track represent a single weekly Top 200 snapshot rather than a peak, cumulative, or average measure of popularity, and the long tail of less-streamed releases that never entered the Top 200 is truncated by construction; the within-chart streaming exponents and Gini coefficients reported in Section 4.6 should therefore be interpreted as descriptors of internal Top 200 concentration rather than as evidence of a population-level power-law popularity distribution. Second, the BERT model used for sentiment analysis was fine-tuned on product reviews rather than song lyrics. To quantify the resulting domain-transfer cost, we externally validated the model against the MoodyLyrics gold standard [31] (Section 4.3.2); BERT achieved 71.5% accuracy, AUC $= 0.795$, and a Spearman rank correlation of $\rho = 0.509$ with the four-class mood labels, with the per-quadrant score ordering (happy $>$ relaxed $>$ angry $>$ sad) matching the expected valence ranking. These values indicate that BERT's signal is genuinely informative on lyric data, while leaving room for improvement via a domain-adapted classifier. Third, BERT's 512-token input limit meant that the primary sentiment scores reflected only the opening window of each track, and 65.8% of analyzed lyrics exceeded this limit (with substantially higher truncation rates for Spanish, 83.6%, and German, 82.7%, than for English, 59.6%, given that Spanish and German lyrics in the corpus tended to be longer). To quantify the resulting bias, we re-scored every track by averaging BERT composites across all $\leq$500-token chunks of the full lyrics (Section 4.3.1); the chunk-averaged and opening-window scores correlated strongly ($r = 0.902$ overall; $r = 0.86$–0.91 within each language) and agreed on positive/negative valence for 88.3% of tracks. Crucially, the by-language means shifted by less than 0.01 in all three languages, leaving the cross-linguistic comparisons reported above qualitatively unchanged. We retain the opening-window scores as the primary measurement, but note that German lyrics, which had the highest median token count (663) and the highest sign-flip rate (17.3%), are the most sensitive to truncation, and that more sophisticated long-document approaches (e.g., hierarchical pooling, Longformer-style attention) would be a useful direction for future work. Fourth, the German bootstrap subsample ($n = 225$) constitutes the full German subcorpus, meaning that the narrow variance ($SD = 0.108$) reflects minimal rather than zero sampling variability. Finally, the box-counting fractal dimension of the rank–frequency scatter ($D \approx 0.92$–0.94) is geometrically trivial given the near-linear Zipf relationship and does not add independent information.

https://doi.org/10.3390/computers15050315

Future work should expand the linguistic coverage to typologically diverse languages (e.g., tonal languages, agglutinative languages), incorporate larger and temporally balanced corpora, and explore domain-adapted sentiment models fine-tuned specifically on song lyrics. The integration of audio features with textual complexity measures represents another promising direction for understanding the multimodal structure of popular music.

6. Conclusions

This study demonstrates that commercially successful song lyrics across English, Spanish, and German share fundamental statistical regularities—Zipfian word distributions, persistent long-range temporal correlations, and multifractal complexity—while differing primarily in sentiment valence and lexical diversity. The most consequential finding is methodological: the dramatic cross-linguistic sentiment disparities reported by English-centric lexicons such as VADER are largely measurement artifacts that disappear under multilingual BERT, reducing a 1.003-point gap to 0.127 points. Within the three Indo-European languages examined—and after controlling for unequal corpus sizes through bootstrap subsampling—multifractal scaling is statistically indistinguishable across English, Spanish, and German, consistent with the interpretation that shared compositional constraints of songwriting impose similar long-range structure. We do not claim universality on the basis of three closely related languages with uneven sample sizes; whether the same patterns generalize to typologically distant languages, non-chart popular music, or larger and more balanced corpora is an important question for future work. These results highlight the importance of language-equitable NLP tools in cross-linguistic research and extend the application of complexity science methods from literary prose to the constrained register of popular song lyrics.

Author Contributions: Conceptualization, F.K., Z.S. and S.B.; methodology, F.K., Z.S. and S.B.; software, F.K. and S.B.; validation, Z.S., S.B. and F.F.; formal analysis, F.K.; investigation, F.K. and S.B.; data curation, F.K.; writing—original draft preparation, F.K. and S.B.; writing—review and editing, F.K., Z.S., S.B., F.F. and N.B.; visualization, F.K.; supervision, Z.S. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Spotify chart data were collected from publicly available Wayback Machine archives. Lyrics were retrieved via the LRCLIB and Genius APIs. The analytical code and derived non-lyric feature tables are available from the corresponding authors upon reasonable request, subject to copyright and data-use restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fell, M.; Sporleder, C. Lyrics-based analysis and classification of music. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 620–631.
2. Shahbazi, Z.; Behnamian, S.; Shahbazi, Z.; Jafari, S. Self-Consistency-Based Fake Media Detection Using Multi-Perspective LLM Reasoning. *Electronics* **2026**, *15*, 1822. [CrossRef]
3. Hu, X.; Downie, J.S. Improving mood classification in music digital libraries by combining lyrics and audio. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, Gold Coast, Australia, 21–25 June 2010; pp. 159–168.
4. Kosmidis, K.; Kalampokis, A.; Argyrakis, P. Language time series analysis. *Phys. A Stat. Mech. Its Appl.* **2006**, *370*, 808–816. [CrossRef]
5. Shahbazi, Z.; Jalali, R.; Shahbazi, Z. Enhancing recommendation systems with real-time adaptive learning and multi-domain knowledge graphs. *Big Data Cogn. Comput.* **2025**, *9*, 124. [CrossRef]
6. Kim, Y.E.; Schmidt, E.M.; Migneco, R.; Morton, B.G.; Richardson, P.; Scott, J.; Speck, J.A.; Turnbull, D. Music emotion recognition: A state of the art review. In Proceedings of the ISMIR, Utrecht, The Netherlands, 9–13 August 2010; pp. 937–952.

7. Laurier, C.; Grivolla, J.; Herrera, P. Multimodal music mood classification using audio and lyrics. In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications*; IEEE: Piscataway, NJ, USA, 2008; pp. 688–693. [CrossRef]
8. Hutto, C.; Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
10. Rogers, A.; Kovaleva, O.; Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [CrossRef]
11. Shahbazi, Z.; Byun, Y.C. Analysis of the security and reliability of cryptocurrency systems using knowledge discovery and machine learning methods. *Sensors* **2022**, *22*, 9083. [CrossRef]
12. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Reading, MA, USA, 1949.
13. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef] [PubMed]
14. Shahbazi, Z.; Byun, Y.C. NLP-based digital forensic analysis for online social network based on system security. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7027. [CrossRef] [PubMed]
15. Williams, J.R.; Lessard, P.R.; Desu, S.; Clark, E.M.; Bagrow, J.P.; Danforth, C.M.; Dodds, P.S. Zipf's law holds for phrases, not words. *Sci. Rep.* **2015**, *5*, 12209. [CrossRef]
16. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689. [CrossRef] [PubMed]
17. Kantelhardt, J.W.; Zschiegner, S.A.; Koscielny-Bunde, E.; Havlin, S.; Bunde, A.; Stanley, H.E. Multifractal detrended fluctuation analysis of nonstationary time series. *Phys. A Stat. Mech. Its Appl.* **2002**, *316*, 87–114. [CrossRef]
18. Montemurro, M.A.; Pury, P.A. Long-range fractal correlations in literary corpora. *Fractals* **2002**, *10*, 451–461. [CrossRef]
19. Drożdż, S.; Oświęcimka, P.; Kulig, A.; Kwapień, J.; Bazarnik, K.; Grabska-Gradzińska, I.; Rybicki, J.; Stanuszek, M. Quantifying origin and character of long-range correlations in narrative texts. *Inf. Sci.* **2016**, *331*, 32–44. [CrossRef]
20. Grabska-Gradzińska, I.; Kulig, A.; Kwapień, J.; Oświęcimka, P.; Drożdż, S. Multifractal analysis of sentence lengths in English literary texts. *arXiv* **2012**, arXiv:1212.3171. [CrossRef]
21. Corral, Á.; Boleda, G.; Ferrer-i Cancho, R. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef]
22. Perotti, J.I.; Billoni, O.V. On the emergence of Zipf's law in music. *Phys. A Stat. Mech. Its Appl.* **2020**, *549*, 124309. [CrossRef]
23. Altmann, E.G.; Pierrehumbert, J.B.; Motter, A.E. Niche as a determinant of word fate in online groups. *PLoS ONE* **2011**, *6*, e19009. [CrossRef] [PubMed]
24. Goel, S.; Watts, D.J.; Goldstein, D.G. The structure of online diffusion networks. In Proceedings of the 13th ACM Conference on Electronic Commerce, Valencia, Spain, 4–8 June 2012; pp. 623–638.
25. Napier, K.; Shamir, L. Quantitative sentiment analysis of lyrics in popular music. *J. Pop. Music Stud.* **2018**, *30*, 161–176. [CrossRef]
26. Hong, Y.; Xue, Y. Emotion classification through song lyrics in multi-languages with BERT. *Appl. Comput. Eng.* **2025**, *108*, 59–68. [CrossRef]
27. De Santis, E.; De Santis, G.; Rizzi, A. Multifractal characterization of texts for pattern recognition: On the complexity of morphological structures in modern and ancient languages. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9833–9846. [CrossRef]
28. Tanaka-Ishii, K.; Takahashi, S. A comparison of two fluctuation analyses for natural language clustering phenomena: Taylor and Ebeling & Neiman methods. *Fractals* **2021**, *29*, 2150033. [CrossRef]
29. Gillet, J.; Ausloos, M. A comparison of natural (English) and artificial (Esperanto) languages: A multifractal method based analysis. *arXiv* **2008**, arXiv:0801.2510. [CrossRef]
30. Nakatani, S. Language Detection Library (Langdetect). 2010. Available online: https://github.com/Mimino666/langdetect (accessed on 8 May 2026).
31. Çano, E.; Morisio, M. MoodyLyrics: A Sentiment Annotated Lyrics Dataset. In Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI 2017), Hong Kong, China, 25–27 March 2017; pp. 118–124. [CrossRef]
32. Alstott, J.; Bullmore, E.; Plenz, D. powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE* **2014**, *9*, e85777. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.