Analysis of Popular Croatian Song Lyrics

Dalibor Bužić
College for Information Technologies
Klaićeva 7, 10 000 Zagreb, Croatia
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varazdin, Croatia
dalibor.buzic@vsite.hr

Jasminka Dobša
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varazdin, Croatia
jasminka.dobsa@foi.hr

Abstract. *The choice of vocabulary, lyric length, and the complexity of the phrasing in popular music songs is shaped by target audience, genre, and thematic focus. Unlike manual methods, by using text mining techniques we can successfully analyse the lyrics of a large number of songs.*
In this study we assembled a data set of 1,188 song lyrics drawn from seventeen well-known Croatian artists. We analysed the differences between individual artists with regard to the length of the text expressed in the number of letters and words, the average length of the word and the diversity of the words used in the songs. Furthermore, we assessed the artists' use of personal pronouns to determine whether they more frequently refer to themselves or address the second person. Characteristic terms for each performer were extracted via TF–IDF, and, finally, latent Dirichlet allocation (LDA) was employed to cluster artists on the basis of thematic content.

Keywords. lyrics analysis, Croatian artists, latent Dirichlet allocation, TF–IDF

1 Introduction

Song lyrics can be regarded as a distinct textual genre that exhibits characteristics of both written and spoken discourse (Schneider, 2020). The aphorism often attributed to Voltaire "Everything too silly to be spoken is sung" humorously underscores the uniqueness of texts created for vocal performance. When enriched by melodic accompaniment and vocal interpretation, song lyrics transcend the constraints that limit poetry intended solely for reading.

Song lyrics are written for a target audience within the musical genre of the performer who will sing them. Some artists remain within a single genre, whereas others range across several; over the course of a career, an artist may even shift genres for various reasons. In addition to these genre-related factors, thematic orientation also drives differences among artists' lyrics: some focus predominantly on love, some on social themes, and some on interpersonal relationships (e.g., family or friendship).

Audience, genre, and theme collectively influence word choice, lyric length, and complexity of expression. Analysing these elements would be difficult to do manually, especially if one wanted to cover a larger number of artists and a larger number of their songs. Therefore, the tasks of data collection, preprocessing, transformation and analysis are delegated to computational text-mining techniques.

Text mining is the process of extracting meaningful information and knowledge from unstructured text (Hotho, Nürnberger & Paaß, 2005). Song lyrics are certainly unstructured: they often lack explicit markup such as chorus labels, sometimes omit punctuation, and the Internet transcriptions seldom originate from the original songwriters. User-generated transcriptions introduce spelling errors or even incorrect words. Heterogeneous contributors generate inconsistencies. For example, some transcribe the chorus only once while others repeat it as many times as it is sung. Extended vowels may be written with repeated characters, and meta-tags such as *Chorus:* or *2×* (indicating a repeated verse) appear sporadically. The names of the artists sometimes appear next to the parts sung in duets. Croatian-language lyrics found online sometimes appear with, and sometimes without diacritics, further complicating preprocessing. Therefore, such specifics must be addressed carefully to ensure robust analysis.

Most lyric-analysis studies focus on English-language songs. Because we found no prior work examining a substantial corpus spanning multiple Croatian artists, our study seeks to fill that gap.

The remainder of this article is organized as follows. In Section 2 we describe related research on lyric analysis. Section 3 details the construction of the Croatian lyrics data set and briefly describes the techniques and methods applied in the research. In Section 4 we present and discuss the results, and in

conclusion we highlight the main findings and outline possible directions for future research.

2 Related research

In (Parada-Cabaleiro et al., 2024), lyrics of popular songs in English were analysed. From a corpus of 353,320 tracks released between 1970 and 2020 the authors extracted lexical, linguistic, and structural descriptors, together with several textual complexity indices. The results revealed that pop lyrics have become simpler and easier to understand over time. In addition, the texts have grown increasingly personal, while the expressed emotions have trended towards greater negativity over time.

Lyric analysis and genre classification were also the focus of the (Fell & Sporleder, 2014). The authors trained a classifier on eight genres (blues, rap, metal, folk, R&B, reggae, country, and religious music) deliberately omitting pop-rock because of its heterogeneity and numerous sub-genres. Using Support Vector Machines, they achieved the highest F1 score for rap (77.6 %), whereas folk proved the most difficult to detect (F1 score of 29.6%). Folk songs were frequently misclassified as blues or country. Classification according to genre was also carried out with eleven human annotators who were less successful than automatic approach.

In (Schedl, 2019) lyric length (in characters and words), repetitiveness, and readability were examined. Significant genre-specific differences emerged in both repetitiveness and readability. Lyrics by very popular artists were highly repetitive but not necessarily easy to read. Among 56 genres, dance and electropop displayed the greatest repetitiveness, while death metal and hardcore punk the least. The lowest readability scores were found for hip-hop, death metal, and progressive metal; the highest for post-rock, psychedelic rock, and dream pop.

In (Hunke, Huber & Steffens, 2025) authors traced the historical evolution of thematic content in German popular songs released between 1954 and 2022. Working with slightly more than 3,000 songs translated into English to leverage the wider range of analytical tools, they applied latent Dirichlet allocation and obtained a six-topic solution comprising (i) music, dance and partying, (ii) society and status, (iii) love and relationships, (iv) desire and self-realisation, (v) dreams and longings, and (vi) miscellaneous. The dominant topic was love and relationships (39 % of all songs). Dreams and longings were prevalent until the mid-1960s, after which their share declined sharply, whereas society and status emerged as the leading topic from 2017 onward. The study also documented a long-term decrease in positive sentiment.

The top-ten songs in the United States (Billboard Hot 100 year-end chart) for each year from 1980 to 2007 were analysed in (DeWall et al., 2011). The focus was on first-person singular and plural pronouns, words signalling social interaction (e.g., *mate, talk*), anger and antisocial behaviour (e.g., *hate, kill*), and positive emotion (e.g., *love, sweet*). Over time, self-focused and antisocial terms increased, while words linked to social interaction and positive emotion declined.

A much longer span, 1970 to 2019, was covered in the study of the Japanese Top 100 singles (Masui & Miyamoto, 2025). Analysis of words that convey emotion revealed a trend of increasing anxiety and decreasing sadness. Cross-correlation and Granger causality tests further showed that increases in disaster related fatalities predicted subsequent rises in positive emotion terms.

In (Rosebaugh & Shamir, 2022) the Universal Data Analysis of Text (UDAT) platform was employed to perform a quantitative study of 18,577 songs by 89 artists, each with at least 100 songs and spanning genres such as rock, pop, hip-hop, soul, R&B, and heavy metal. Automatic genre classification achieved 12.3 % accuracy (random guess is ~1.1 %). Readability indices differed markedly across artists, with hip-hop and rap exhibiting the most complex lyrics. Further analyses of sentiment, gender-identity markers, and lexical diversity showed that lyrics expressing positive sentiment were generally easier to read, whereas less positive songs tended to be linguistically more complex.

A comparative evaluation of traditional (LDA) and transformer-based approaches (BERT, GPT, and XLNet) to topic modeling across two datasets was conducted in (Riaz et al., 2025). Using topic-coherence and topic-diversity metrics, a hybrid XLNet-BERT model achieved the best overall performance, outperforming LDA, standalone BERT, and the GPT-LDA, XLNet-LDA, and GPT-BERT combinations.

As most lyric analyses focus on English-language song lyrics, we employed text-mining methods and techniques to investigate songs by popular Croatian artists - a comparatively underexplored domain. To make such an inquiry feasible, we the first step was to compile a corpus of song lyrics.

3 Methods

This section describes the methodology of analysis of Croatian song lyric. The first subsection describes creation of data set of Croatian song lyrics, while the second and the third subsection briefly describe the TF-IDF weighting scheme and LDA method used.

3.1 Data set

The data set comprises song lyrics by 17 Croatian music artists, obtained via web scraping. An artist was included if at least 40 songs could be collected, while

a modest degree of genre diversity was maintained. Each lyric file contains three attributes: artist, song title, and lyrics. All songs with fewer than 50 words were checked manually; those with truncated or missing texts (about one hundred songs) were discarded. Using the *textcat* package in R programming language, we removed lyrics not written in Croatian (six in English, two in German, and one each in Latin and Slovene), yielding a final corpus of 1,288 songs. Fig. 1 lists the selected artists together with the number of songs contributed by each.

Several instances were noted in which a song was attributed to a performer other than the original singer. For example, *Zagrli me*, credited to Arsen Dedić even though Zdravko Čolić was the first to record it (Dedić authored the song and occasionally performed it). Such cases were retained.

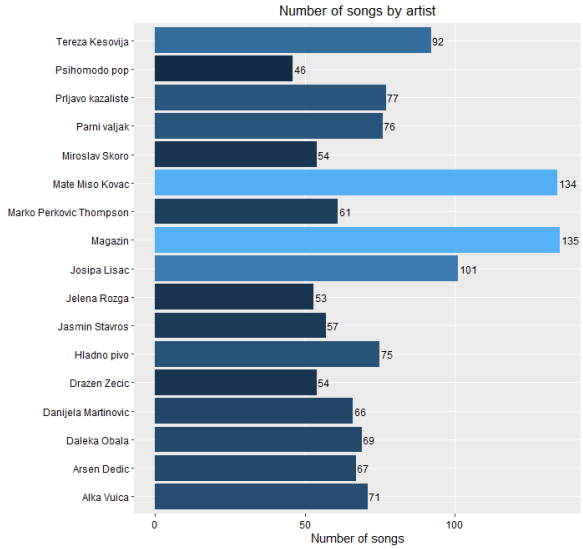

Figure 1 Number of songs by artist

The corpus displayed multiple inconsistencies requiring corrective preprocessing. About twenty texts contained metadata that were not part of the actual lyrics; these segments were removed manually. Some words featured repeated letters indicating elongated vocal delivery; such sequences were programmatically collapsed to single characters (e.g., *jeeeedaaan žiiiivoooot iiimaaamooo* to *jedan život imamo*). Because some song lyrics were written with Croatian diacritics (č, ž, š …), whereas the remainder did not (c, z, s …), all diacritics were normalised to their non-diacritic counterparts.

Stanzas are typically not repeated verbatim in a transcription, yet some files recorded every occurrence of a sung stanza. For instance, *Ovo nije moje vrijeme* by Daleka Obala appeared with eight four-line stanzas, although, without repetition, it contains only three. Duplicate stanzas were therefore removed programmatically (affecting 350 songs), and all subsequent analyses were performed on the de-duplicated texts. Repeated lines within a single stanza were preserved.

3.2 TF-IDF

Term frequency–inverse document frequency (TF–IDF) is a weighting scheme composed of two factors: term frequency (TF) and inverse document frequency (IDF). TF is the proportion of occurrences of a term relative to the total number of terms in the document; in the phrase *starry starry night*, for example, TF(*starry*) = 0.667 and TF(*night*) = 0.333. IDF is computed as

$$\mathrm{IDF} = \log(\mathrm{N}/\mathrm{df}(\mathrm{t})) \quad (1)$$

where N is the number of documents in the corpus and $df(t)$ the number of documents containing term t. Rare terms therefore receive high IDF values, whereas frequent terms receive low ones (Manning, 2009, pp. 117-118). The product TF × IDF highlights those words that are most characteristic or distinctive of a given document within the entire collection.

Owing to its simplicity and effectiveness, TF–IDF is frequently used in text mining tasks, including lyric-related research. In (Van Zaanen & Kanters, 2010) TF-IDF was employed to construct a mood-based lyric-classification system; in (Sahera & Ilahi, 2025) it was used for lyric-driven song retrieval within Spotify; in (Hori, 2019) TF-IDF was combined with *word2vec* to extract colour terms from lyrics, supplying inputs to an image-generation pipeline; and in (Schneider, 2020) it was demonstrated how TF-IDF can be exploited to identify signature words of particular performers, periods, or albums.

3.3 LDA

Latent Dirichlet allocation (LDA) is one of the most widely used topic-modelling techniques. It is a generative probabilistic model that represents each topic as a distribution over words and each document as a mixture of those topics (Jelodar et al., 2019). In essence, LDA assumes that documents contain multiple latent themes; by analysing word co-occurrence patterns, the algorithm infers both the topics and the degree to which each document belongs to them.

LDA has been applied to lyrics for various purposes. For example, in (Sasaki et al., 2014) LDA was integrated into an interactive lyric-search system that visualises topics for exploratory queries. In (Dakshina & Sridhar, 2014) LDA was used to facilitate emotion recognition in song texts, while in (Laoh, Surjandari & Febirautami, 2018) it was employed to model topics in Indonesian lyrics.

4 Results and Discussion

The shortest song in the corpus contains only 14 words, the longest 329, and the mean lyric length is 103.6 words. Fig. 2 displays the distribution

of word counts per song. Lemmatization and stemming were not used, nor were stop words removed.

Fig. 3 lists the longest single lexical items encountered. The two lengthiest forms, *oblibabuibajeijeije* and *trajnaninananana*, are semantically nonsensical, a phenomenon not uncommon in lyrics.

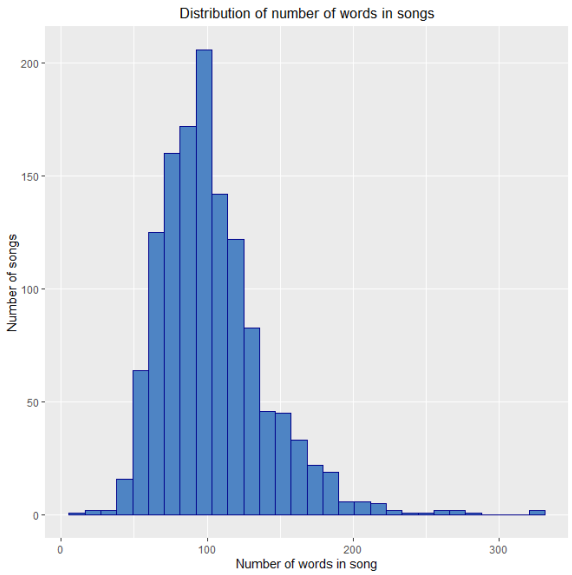

Figure 2 Distribution of number of words in songs

Figure 3 The longest words

Table 1 summarises lyric characteristics for all artists under study. Within each column, the three highest values are highlighted in shades of green and the three lowest in shades of blue.

The average character count for Hladno Pivo exceeds that of Mate Mišo Kovač, the artist with the briefest texts, by 45 %. Across the corpus, the mean number of words (not unique) is 103.6 per song; lexical richness varies by 40.9 % between the most vocabulary-sparse artist (Psihomodo Pop) and the richest (Arsen Dedić). Hladno Pivo also exhibits the greatest mean word length (4.39 characters), followed by Psihomodo Pop and Marko Perković Thompson. The shortest average word lengths occur in the lyrics of Jelena Rozga, closely followed by Magazin and Danijela Martinović.

Table 1 Average lyrics statistics per artist

Artist	Letters	Words	Word length	Diversity
Alka Vuica	462.7	116.6	3.96	0.29
Arsen Dedic	477.3	119.2	3.99	0.28
Daleka Obala	397.0	101.4	3.96	0.28
Danijela Martinovic	402.2	103.0	3.90	0.27
Drazen Zecic	366.3	93.4	3.95	0.26
Hladno pivo	492.7	113.5	4.39	0.39
Jasmin Stavros	440.4	112.0	3.94	0.24
Jelena Rozga	455.4	117.2	3.88	0.28
Josipa Lisac	424.4	108.4	3.97	0.23
Magazin	409.2	105.6	3.88	0.21
Marko Perkovic Thompson	417.7	99.8	4.19	0.34
Mate Miso Kovac	339.8	85.8	3.97	0.21
Miroslav Skoro	374.9	93.8	4.00	0.36
Parni valjak	423.9	104.1	4.09	0.26
Prljavo kazaliste	370.7	91.4	4.07	0.27
Psihomodo pop	357.6	84.6	4.24	0.35
Tereza Kesovija	446.9	111.7	4.00	0.24
Averages	**415.2**	**103.6**	**4.02**	**0.28**

The final column reports average lexical diversity (the ratio of the number of unique words to the total number of words in the song). Higher values indicate fewer repetitions within a song. Hladno Pivo shows the greatest diversity, followed by Miroslav Škoro and Psihomodo Pop; the lowest diversity is found in the repertoires of Mate Mišo Kovač, Magazin, and Josipa Lisac.

At the song level, four songs exhibited a diversity score below 0.30, with word counts ranging from 26 to 117. Conversely, twenty-three songs achieved diversity above 0.90, their lengths varying between 83 and 217 words.

Next, we examined inter-artist differences in the use of the first-person singular – *ja* (I), second-person singular – *ti* (you), and first-person plural – *mi* (we) pronouns. In Croatian, personal pronouns inflect for seven cases, and some cases have both long and short forms (Težak & Babić, 1992, pp. 106-108). Several inflected forms are shared across cases, yielding the following forms:

- ja: *ja, mene, me, meni, mi, mnom, mnome*
- ti: *ti, tebe, te, tebi, tobom*
- mi: *mi, nas, nama, nam*

A complication arises because *mi* is ambiguous: it serves as the dative of *ja* (I) and as the nominative

plural of *mi* (we). To avoid conflating the two, every token *mi* was counted exclusively as the plural nominative.

Table 2 Relative frequencies of pronouns

Performer	Ja (I) (%)	Ti (you) (%)	Mi (we) (%)
Alka Vuica	4.30	4.14	1.85
Arsen Dedic	2.07	2.44	1.68
Daleka Obala	3.56	2.40	1.99
Danijela Martinovic	3.81	4.26	2.35
Drazen Zecic	3.75	3.41	1.80
Hladno pivo	1.87	1.48	1.69
Jasmin Stavros	3.84	3.21	1.79
Jelena Rozga	3.53	3.75	1.96
Josipa Lisac	2.61	3.01	1.53
Magazin	4.09	3.50	2.36
Marko Perkovic Thompson	2.60	2.45	1.97
Mate Miso Kovac	2.97	3.63	1.94
Miroslav Skoro	3.36	2.37	1.84
Parni valjak	3.72	1.90	1.86
Prljavo kazaliste	3.51	2.50	1.51
Psihomodo pop	3.88	2.16	1.34
Tereza Kesovija	2.59	2.93	1.77
Averages	**3.26**	**2.97**	**1.87**

Table 2 reports the relative frequencies of these pronouns. Values represent the ratio of pronoun tokens to the total word count in each artist's lyrics. On average, *ja* accounts for 3.26 % of all words; the highest proportion occurs in songs by Alka Vuica (4.30 %), whereas Hladno Pivo uses *ja* 2.3 times less frequently. The second-person *ti* appears slightly less often overall (2.97 %). Danijela Martinović addresses the listener most intensively, with *ti* comprising 4.26 % of her lyrics. Alka Vuica's repertoire combines the greatest aggregate share of *ja* + *ti* (8.44 %), while Hladno Pivo registers the lowest (3.35 %). References to *mi* are led by Magazin (2.36 %), closely followed by Danijela Martinović (2.35 %), whose texts contain the highest cumulative proportion (10.43 %) of all three pronoun categories. The largest *ja*–*ti* imbalance in favour of *ja* is exhibited by Parni Valjak (1.82 %), with Psihomodo Pop a close second (1.72 %). On the opposite side, Mate Mišo Kovač shows a negative difference (-0.66 %), indicating a stronger orientation toward the listener (*ti*) than toward the self (*ja*).

Fig. 4 displays, for each performer, the ten highest-scoring TF–IDF terms. In constructing the TF–IDF matrix we removed 192 Croatian stop words. It is interesting how TF-IDF gave importance to the words that denote the topic the performer is often singing about. For example, Miroslav Škoro's repertoire, frequently centred on rural life in eastern Croatia, is marked by *konji* (horses), *šor* (village lane), *Dunav* (Danube), *bećar* (reveler), and *bake* (grandmothers). Among all artists examined, Mate Mišo Kovač sings most consistently about Dalmatia and nostalgia; TF–IDF surfaces the words *stijena* (rock), *mama* (mother), *Dalmacija*, *Hrvati* (Croats), and *jata* (flocks). Marko Perković Thompson's lyrics are largely patriotic and martial, signalled by *sloboda* (freedom), *krv* (blood), *narod* (nation), *Hrvati*, *braniti* (defend), and *ginuti* (die). Dražen Zečić often addresses love problems, reflected in *voljeti* (to love), *slomiti* (break), *pokidati* (tear apart), *objasniti* (explain), and *isprika* (apology). By contrast, performers with a more diffuse thematic palette (such as Arsen Dedić,

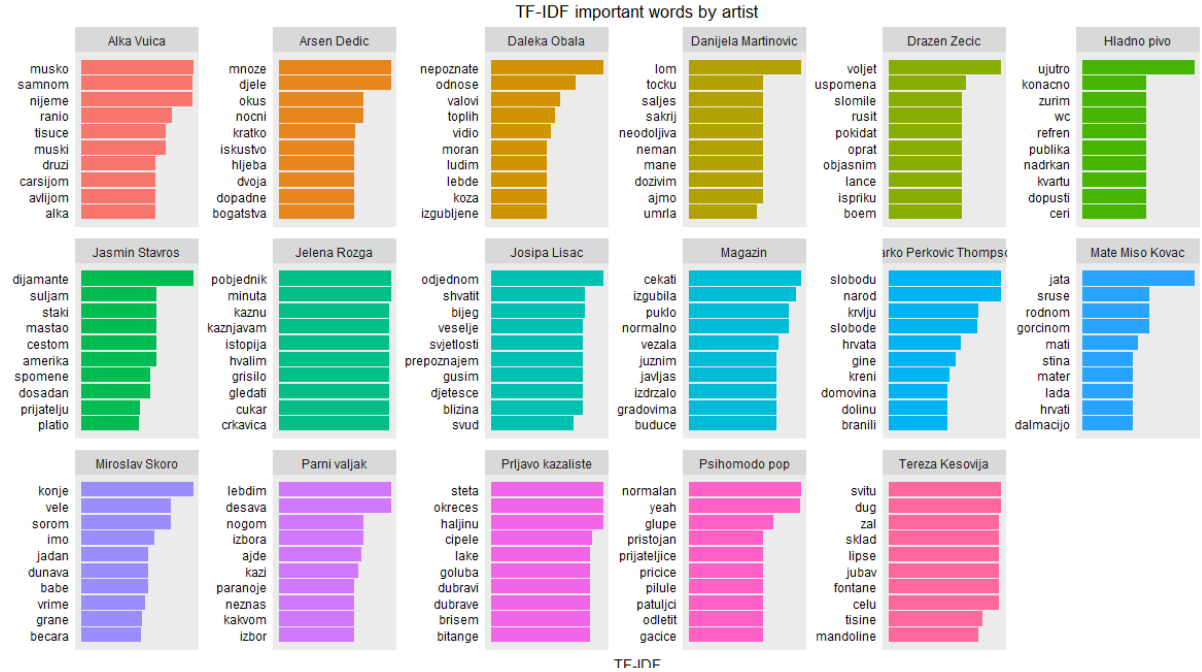

Figure 4 Top TF–IDF terms per artist

Hladno Pivo, and Josipa Lisac) exhibit TF–IDF term sets whose connection to a single topic is less readily discernible.

Using the LDA() function from the R package *topicmodels* with Gibbs sampling, we estimated a seven-topic model. The number of topics was determined qualitatively – by inspecting the highest-probability words within each candidate topic. Table 3 lists, for each topic, the ten words with the highest posterior probability. Topic 1 is dominated by *srce* (heart), *duša* (soul), *žena* (woman), *čovjek* (man), and *ljudi* (people). Topic 2 is characterised by *noć* (night), *dan* (day), *san* (dream), *put* (road), and *kraj* (end).

Table 3 Top 10 words per topic

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
srce	noc	uvijek	sad	ima	vise	noci
zena	dan	nikad	znam	dobro	zna	ljubav
dusa	kraj	ljubavi	zivot	zivot	duso	oci
jednom	put	opet	mogu	treba	vidim	ruke
nocas	san	znas	sada	vrijeme	tko	more
dva	jedan	rijeci	nista	dana	zato	mora
prije	mom	ljubav	vise	malo	ljubav	ime
ljudi	znam	zasto	volim	vrata	dao	sunce
covjek	glas	zbog	nekad	nemam	druge	gledam
tuga	bit	zelim	mozda	ovo	svijet	sva

Finally, for each song, we extracted the probabilities of belonging to each of the seven topics from the LDA model. We then aggregated these probabilities by performer by summing, across all songs by a given artist, the topic-membership probabilities. For each performer, we identified the largest of the seven aggregate probabilities and assigned the artist to the corresponding topic. For example, Josipa Lisac obtained the following aggregates: 13.1 (Topic 1), 17.6 (Topic 2), 14.5 (Topic 3), 14.6 (Topic 4), 13.7 (Topic 5), 13.8 (Topic 6), and 13.7 (Topic 7). The highest value occurred for Topic 2; accordingly, Josipa Lisac was assigned to that topic. The results are presented in Fig. 5.

To aid the interpretation of our results, we drew on the artists' genre affiliations. Genre information was extracted from each artist's Wikipedia page. Some artists are linked to a single genre, others to several; the record holder is Parni Valjak, labelled with six genres. Miroslav Škoro is the sole exception - his Wikipedia entry lists no genre, so we assigned him to tamburitza music. In several cases the Croatian- and English-language Wikipedia pages disagree on artist's genre, some performers briefly shifted genres during their careers, and certain categories either overlap (e.g., pop vs. popular (*zabavna*) music) or subsume one another (pop-rock overlaps both pop and rock).
To reduce this complexity, we merged closely related labels (e.g., pop and *zabavna* into *pop*) and limited each artist to at most two categories - compressing, for instance, "pop," "pop-rock," and "rock" into a single tag (*pop-rock*). The simplified genre assignments are listed in Table 4. Three artists (Tereza Kesovija, Josipa Lisac and Marko Perković Thompson) are associated with two genres; the remaining performers are allocated to one. Overall, pop is the most prevalent category, encompassing ten of the seventeen artists.

Table 4 Genre assignments

Genre	Artist
pop	Jasmin Stavros, Magazin, Jelena Rozga, Mate Mišo Kovač, Dražen Zečić, Danijela Martinović, Alka Vuica, Tereza Kesovija, Josipa Lisac, Marko Perković Thompson
pop-rock	Prljavo kazalište, Parni valjak
rock	Daleka obala, Josipa Lisac
punk-rock	Hladno pivo, Psihomodo pop
chanson	Arsen Dedić, Tereza Kesovija
tamburitza	Miroslav Škoro
Christian rock	Marko Perković Thompson

In Topic 1, the model groups three pop performers. Topic 2 contains Josipa Lisac and genre-compatible Daleka Obala, but also Tereza Kesovija, whose placement is less intuitive; nonetheless, similar thematic material can cross genre lines, so lyrics alone may not suffice for classification. Alka Vuica stands alone in Topic 3; while nominally a pop artist, her lyrics diverge markedly from other pop artists. Miroslav Škoro is the only tamburitza representative, and many of his songs address themes absent from the rest of the corpus. Topic 5 clusters the stylistically related punk-rock bands Psihomodo Pop and Hladno Pivo, though the inclusion of pop music singer Jasmin Stavros appears

Figure 5 Artists grouped per topic

anomalous. Topic 6 unites the pop rock groups Parni Valjak and Prljavo Kazalište - an entirely reasonable grouping. Topic 7 contains the largest number of artists; in our judgment, Arsen Dedić does not belong here, whereas Magazin would fit better alongside former lead vocalists Jelena Rozga and Danijela Martinović. Although Marko Perković Thompson records some pop material, we would, like Miroslav Škoro, probably place him in his own topic.

Taken as a whole, the LDA model groups artists reasonably well from a genre perspective, especially given that identical or similar themes recur across multiple genres and that artists are not strictly bound to a single style (despite our coarse reduction of both artists and genres to one or two principal label).

5 Conclusion

In this study we examined the lyrics of 1,188 songs by 17 Croatian artists, with 46 to 135 songs per artist. On average a song contained around 104 words and 415 characters, and the mean word length was 4.02 characters. The longest texts were produced by Hladno Pivo - 45 % longer than those of Mate Mišo Kovač, who had the shortest. Arsen Dedić's songs employed 41 % more tokens than those of Psihomodo Pop, the artist with the fewest.

Lexical diversity, defined as the ratio of unique words to total number of words, ranged from 0.21 for the pop singer Mate Mišo Kovač to 0.39 for the punk-rock group Hladno Pivo, a relative difference of 62 %. Across the entire corpus, first-person singular forms (*ja*) accounted for 3.26 % of all words and second-person singular forms (*ti*) for 2.97 %, indicating only a modest preference for self-reference.

A TF-IDF matrix yielded the ten most distinctive words for each artist. For performers whose thematic focus differs markedly from the rest (e.g., Mate Mišo Kovač, Marko Perković Thompson, Dražen Zečić), TF-IDF surfaced terms that distinguish their textual expression from others. However, for artists with more dispersed themes that intertwine with other artists, the words extracted by TF-IDF are more difficult to associate with that artist.

Using latent Dirichlet allocation (LDA) we induced seven topics. Lyrics were aggregated by artist, and each artist was assigned to one topic. We independently associated every artist with one or at most two genres drawn from a seven-category scheme. We looked at the artists in each theme from the aspect of the musical genre to which the artist belongs. Genre-wise, LDA has satisfactorily classified the performers into groups (themes) in this way, especially when considering that song lyrics often talk about the same or similar themes.

The present data set is dominated by pop lyrics. Extending it with additional genres would yield a more balanced dataset, enabling genre classification experiments that combine LDA with other machine-learning algorithms. It would be particularly interesting to explore contemporary deep-learning approaches, specifically transformer-based methods such as BERTopic or ChatGPT. In addition, the presence of emotions in songs could be investigated and whether there are significant differences in emotions between Croatian artists.

References

Dakshina, K., & Sridhar, R. (2014). LDA based emotion recognition from lyrics. In Advanced Computing, Networking and Informatics-Volume 1: Advanced Computing and Informatics *Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014)* (pp. 187-194). Cham: Springer International Publishing.

DeWall, C. N., Pond Jr, R. S., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular US song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, 5(3), 200.

Fell, M., & Sporleder, C. (2014). Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics*: Technical papers (pp. 620-631).

Hori, G. (2019). Color extraction from lyrics. In *Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering* (pp. 1-6).

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *Journal for Language Technology and Computational Linguistics*, 20(1), 19-62.

Hunke, T., Huber, F., & Steffens, J. (2025). The Evolution of Song Lyrics: An NLP-Based Analysis of Popular Music in Germany from 1954 to 2022. *Music & Science*, 8, 20592043251331155.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78(11), 15169-15211.

Laoh, E., Surjandari, I., & Febirautami, L. R. (2018). Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 270-274). IEEE.

Manning, C. D. (2009). *An introduction to information retrieval*. Cambridge Press.

Masui, H., & Miyamoto, Y. (2025). Emotions in Japanese song lyrics over 50 years: Trajectory over time and the impact of economic hardship and disasters. *Current Research in Ecological and Social Psychology*, 8, 100218.

Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., & Zangerle, E. (2024). Song lyrics have become simpler and more repetitive over the last five decades. *Scientific Reports*, 14(1), 5531.

Riaz, A., Abdulkader, O., Ikram, M. J., & Jan, S. (2025). Exploring topic modelling: a comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity. International Journal of Electrical & Computer Engineering (2088-8708), 15(2).

Rosebaugh, C., & Shamir, L. (2022). Data science approach to compare the lyrics of popular music artists. *Unisia*, 40(1), 1-26.

Sahera, N. J., & Ilahi, S. B. N. (2025). Utilization Of TF-IDF Weighting In Song Search System Based On Spotify Lyrics. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 4(3), 1920-1927.

Sasaki, S., Yoshii, K., Nakano, T., Goto, M., & Morishima, S. (2014). Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. In *Ismir* (pp. 585-590).

Schedl, M. (2019). Genre differences of song lyrics and artist wikis: An analysis of popularity, length, repetitiveness, and readability. In *The World Wide Web Conference* (pp. 3201-3207).

Schneider, R. (2020). A corpus linguistic perspective on contemporary German pop lyrics with the multi-layer annotated “Songkorpus”. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 842-848).

Težak, S., & Babić, S. (1992). *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje*. Zagreb: Školska knjiga.

Van Zaanen, M., & Kanters, P. H. M. (2010). Automatic mood classification using tf* idf based on lyrics. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 75-80). TiCC.