

Article

Music Genre Classification Using Prosodic, Stylistic, Syntactic and Sentiment-Based Features

Erik-Robert Kovacs ^{1,*}  and Stefan Baghiu ^{2,*}¹ Department of Cybernetics, Statistics and Economic Informatics, Bucharest University of Economic Studies, Piața Romană nr. 6, Sector 1, 010374 București, Romania² Department of Romance Studies, Lucian Blaga University of Sibiu, Bd-ul. Victoriei, Nr.10, 550024 Sibiu, Romania

* Correspondence: erik.kovacs@csie.ase.ro (E.-R.K.); stefan.baghiu@ulbsibiu.ro (S.B.)

Abstract

Romanian popular music has had a storied history across the last century and a half. Incorporating different influences at different times, today it boasts a wide range of both autochthonous and imported genres, such as traditional folk music, rock, rap, pop, and manele, to name a few. We aim to trace the linguistic differences between the lyrics of these genres using natural language processing and a computational linguistics approach by studying the prosodic, stylistic, syntactic, and sentiment-based features of each genre. For this purpose, we have crawled a dataset of ~14,000 Romanian songs from publicly available websites along with the user-provided genre labels, and characterized each song and each genre, respectively, with regard to these features, discussing similarities and differences. We improve on existing tools for Romanian language natural language processing by building a lexical analysis library well suited to song lyrics or poetry which encodes a set of 17 linguistic features. In addition, we build lexical analysis tools for profanity-based features and improve the SentiLex sentiment analysis library by manually rebalancing its lexemes to overcome the limitations introduced by it having been machine translated into Romanian. We estimate the accuracy gain using a benchmark Romanian sentiment analysis dataset and register a 25% increase in accuracy over the SentiLex baseline. The contribution is meant to describe the characteristics of the Romanian expression of autochthonous as well as international genres and provide technical support to researchers in natural language processing, musicology or the digital humanities in studying the lyrical content of Romanian music. We have released our data and code for research use.

Keywords: musical lyrics classification; musical genre classification; lexical analysis; syntactic analysis; prosody analysis; sentiment analysis

Academic Editors: Sardar Jaf,
Basel Barakat and Yongqiang Cheng

Received: 20 October 2025

Revised: 6 November 2025

Accepted: 15 November 2025

Published: 19 November 2025

Citation: Kovacs, E.-R.; Baghiu, S. Music Genre Classification Using Prosodic, Stylistic, Syntactic and Sentiment-Based Features. *Big Data Cogn. Comput.* **2025**, *9*, 296. <https://doi.org/10.3390/bdcc9110296>

Copyright: © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Literary, musical, and film genres are all social constructs. Although many conservative art theorists have created the illusion that genre is something intrinsically embedded in the work of art, the one thing we know about genre is that it is more of a convention than a datum. As Franco Fabbri has argued in his “A Theory of Musical Genres,” a musical genre is “a set of musical events (real or possible) governed by a definite set of socially accepted rules” [1]. Fabbri thought that genres had ‘rules,’ and that these rules are to be decided by several factors, including formal and technological ones (composition, instrumentation, etc.) or semiotic ones (relating to meaning and message) [1]. Moreover, the formal and

semiotic rules of genres are doubled by how audiences behave at the performance of music, how society reacts to the said musical genre, or how they are marketed within capitalist structures of production and counterproduction (i.e., the production of needs). However, Fabian Holt, the author of the 2007 monograph *Genre in Popular Music* claimed that the mass consumption of music makes genre extremely difficult to define [2]. Most scholars in the field agree that genres are not pure or exclusive categories, and that a song can be assigned to multiple genres. Of course, in narrative and literary studies, the concepts of multiple genres and microgenres [3,4] have prompted research that seeks to move beyond a canonical or fixed understanding of writing itself. A novel, for instance, can not only belong to multiple genres at once, but its constituent parts may also fall into different microgenres. For Fabbri, since genres are socially convened, the convention is an arbitrary solution that historicizes itself into a fixed or grounded definition [1]. Moreover, with the commercial atomization of genres on music stores and platforms [5], the concept of genre has remained strictly a commercial label for consumers to identify their product. Although it might seem that genres have fragmented and multiplied over time, studies show that they have become more homogeneous as more music has been produced, a process labeled as genre generalization [6]. If we take out of the equation the professionals of music, the niche searchers and those obsessed with specific genres also active online [7], who are mostly still active in the “independent music marketplace” [8], the mass of the public generally assumes that there are only a few “classic” popular genres. However, if one takes a look at the labeling system of an online platform of music distribution, it becomes clear that genres and subgenres are interconnected and interdependent [9].

The main question of this article is if one can recognize music genres from song lyrics. Of course, hearing a hip-hop beat or a rock song can stem instant recognition. But what about reading the text? This exciting perspective, that of being able to read some lyrics and immediately know what kind of music they should be accompanied by, is our motor in engaging this knowledge process. Scholars like Dai Griffiths have argued that lyrics in songs function differently than in poetry [10]. Lyrics are transformed and conceived in relation to the music they are assigned, and the reading process of lyrics differs substantially to reading poetry [10]. There are many ways this question can be studied. Recent approaches to humanities research leverage the wide availability of digital text data to study the stylistic properties of texts using a computational approach [11].

The article aims to use a similar general approach, albeit one reflecting the empirical computational social science methodology [12,13]. We start by gathering and describing a new Romanian song lyrics dataset obtained from a user-contributed website, *tabulaturi.ro*. We use a Romanian-language slur dataset which we have gathered, together with an improved version of the SentiLex sentiment analysis library [14] which we have manually rebalanced and enriched, to define a set of prosodic, stylistic, syntactic, and sentiment-based features which can be used to encode song lyrics in the Romanian language. These features have the advantage of being domain-specific and immediately interpretable. Based on the feature-space representation of the lyrics, we train a logistic regression classification model able to recognize the genre of each song with 62.75% accuracy and an F1 score of 69.94%. The statistical properties of the logistic regression model, as well as feature profiles built for each genre, support our formal analysis of the lyrics in terms of the influence and importance of each feature in predicting the genre, leading to some counter-intuitive conclusions such as revealing deep linguistic similarities between otherwise culturally and historically conflicting genres. We have made the datasets available for research use.

The remainder of the article is organized thusly. In Section 2, a brief summary of current research into music lyrics classification is given. Section 3 outlines the data and the procedure used to train and evaluate the models. Section 4 presents the results obtained

by applying the models to the domain data. Section 5 contains an applied interpretation of the results, focusing on the statistical differences between the genres. Finally, Section 6 provides the conclusions which can be drawn from this research.

2. Literature Review

Music genre classification has been studied in the context of music information retrieval systems [15]. With the growth of digital music content on the Internet [16,17], the importance of automatically classifying music files became increasingly necessary due to the cost in time and resources for labelling by experts or end-users [15]. However, the training data for such labelling has been grounded in the outsourcing of the classification itself first to artists and labels. While studies and corporate experience have shown that automatic classification is crucially improved by “different types of data” [18], Soundcloud, DistroKid, and other platforms which spread music across the Internet have also asked uploaders to select their genre, crowdsourcing this data, at the same time making it difficult for third parties to retrieve it. Initially, great consideration was given to the classification of music based on the actual audio data present in digital audio files, based on the premise that songs within a genre tend to share the same instrumentation, rhythmic patterns and other sonic characteristics. At the same time, genres with similar audio characteristics often get misclassified due to overlapping features (like blues and country or hip-hop and reggae) [19]. For this reason much of the early work on machine learning music genre classification devoted significant effort to the areas of feature extraction and engineering [15], aiming to represent these characteristics accurately and efficiently. In an early contribution, Li et al. use an SVM and signal processing techniques to extract features and improve the accuracy of digital tracks classification by genre [15]. Their general approach, where only the audio files are taken into consideration, came to be known as content-based genre classification. Of late, classification has evolved with the implementation of audio features in genre classification, with a high emphasis on working with Mel-Frequency Cepstral Coefficients and improving through Ensemble Stacking Models [19,20].

It is also possible to use neural networks for content-based genre classification. Multi-Layer Perceptron (MLP) models can be used to assess feature importance across genres using permutation [19]. Pelchat and Gelowitz contribute an interesting approach where songs are first converted into 2-D spectrograms (graphical representations of the distribution of frequencies across time in a sound sample) and then computer vision techniques (specifically a Convolutional Neural Network) are used on slices of each spectrogram for the classification phase [21].

Similarly to instrumentation, rhythm, and other aural characteristics, musical genres can also be distinguished by specific lyrical themes and patterns. Natural language processing techniques can be used on datasets of song lyrics to classify songs using features extracted from the text. Despite significant progress in the NLP field over the last decade, lyrics analysis-based genre classification has been relatively less used when compared to content-based approaches to music genre classification. Tsaptsinos proposes a Recurrent Neural Network architecture for this purpose, noting the importance of choosing appropriate hand-built features demanded by traditional machine learning approaches, which is a significant drawback of these techniques [22]. Deep learning approaches can use embedding vectors instead of these hand-built features, simplifying the feature engineering process. He reports modest accuracy (46%) but with the advantage of using a Hierarchical Attention [23,24] mechanism which assigns weights to different words, making the model more explainable. If contributing to a theory of genre is one of our aims, explainability becomes an important characteristic which we also aim to preserve.

In respect to research on lyrics and poetry, Singhi and Brown have shown that there are important overlapping features of songs and poems, but that the differences between them come from different adjectives to express the same concepts [25]. While song lyrics and poetry verses differ in style, intent, linguistic choices, and function, they can also blur in cases like poetic lyricists. More recent text-based approaches also exist, such as the work of Akalp et al., who used Transformers-based models such as BERT to classify song lyrics, obtaining a 77.63% accuracy [26]. Nevertheless, there is a case to be made for hand-built features. Mayer et al. have used stylistic features to characterize the style and rhyme of song lyrics [27]. One major advantage of these is that they are immediately interpretable; the density and high dimensionality of embedding vectors renders them less so. Mayer et al. use features directly related to the study of prosody, that is, the formal structure of poetic texts, such as the rhyming scheme and vocabulary [27]; thus the classification model's decisions can be interpreted qualitatively too.

Finally, a hybrid, or multimodal approach is also possible, combining features based on the aural characteristics of the songs as well as those based on the lyrics. An early example is Mayer and Rauber's work which uses a multi-component ensemble classifier, with individual classifiers trained on combinations of audio and text features [28]. As part of their text features, they use topic (word choice) in addition to rhyme and style features [28]; whereas for the audio features, they primarily use rhythm-based ones. A more recent example highlighting some of the challenges with this approach is found in Li et al., who use multimodal ensemble methods consisting of a CNN and BERT models to classify the songs with an F1 score of 87% [29].

More recently, Watanabe and Goto have proposed establishing a new field, Lyrics Information Processing (LIP), which should be a distinct interdisciplinary field at the intersection of Natural Language Processing (NLP) and Music Information Retrieval (MIR) [30]. In their argument, they stress out that lyrics have unique characteristics (rhyming, emotional expression, structure) that standard NLP tools don't address well [30]. By encouraging research combining NLP, MIR, HCI, signal processing, and linguistics, they aim for a new approach to music studies, one which might be important for our future research as well. For now, we aim to see if the recognition of music genres can be based on lyrics only.

This is why referring to the work of Computational Stylistics in Poetry, Prose, and Drama is especially relevant. Computational stylistics, by means of metrics, large-scale pattern recognition, and digital analysis, allows scholars to go beyond subjective impressions and examine structural and linguistic features in a systematic way. As highlighted in "The Polite Revolution of Computational Literary Studies" [11], this approach does not aim to replace traditional close reading but rather to enrich it, offering a balanced and collaborative integration of quantitative methods and humanistic interpretation. By applying algorithms to rhythm, meter, rhyme patterns, and lexical distribution, researchers can uncover hidden stylistic consistencies across poems and lyrics, as well as subtle divergences that might be invisible to the naked eye. Such computational methods provide a new perspective on how poetic language functions, enabling both distant reading [31] across large corpora and a more nuanced appreciation of individual texts.

The current paper attempts to fill a gap in the existing research by providing a novel dataset in a lower-resource language, Romanian, while also covering an often-overlooked and trivialized subject, namely popular music; we also contribute tools to support the development of Digital Humanities methodologies for the study of Romanian popular music. In addition to this, within the broader landscape of computational music genre classification literature, starting from the feature-space representations of the lyrics using our encoding as well as the parameters of the classifier, we develop new insights in the theory of genre and support them on prosodic, stylistic, syntactic, and sentiment-based

linguistic features, showing that lyrics are indeed strongly related to the genre, although not in a straightforward way.

3. Data and Methods

We present our dataset, then proceed to describe the labeling and preprocessing process. We then discuss the lexical analysis library improvements we have made to represent the data in a feature space consisting of hand-designed lyrics-specific features. Then, we use classical supervised machine learning algorithms (logistic regression) to train a classifier to predict the genre using these representations.

3.1. Training Data

One of the issues with the machine learning approach to musical genre classification is the lack of access to large enough, high-quality datasets. Tsaptsinos used a proprietary dataset; nevertheless the genre information was not present, so a third-party API was used to annotate it [22]. The issue of training data availability is compounded by the status of the Romanian language as a relatively low-resource language (although current efforts for digitization, etc., will in time mitigate this). To the best of our knowledge there are no currently available song lyrics datasets available in the Romanian language, with or without genre, artist, etc., annotations. As such, compiling our own dataset seemed to be the best option. We used the publicly available, user-contributed (crowdsourced) lyrics available on the *tabulaturi.ro* website (<https://www.tabulaturi.ro/> (accessed on 31 October 2024)) as a starting point.

According to the website itself, *tabulaturi.ro* is the largest online archive of Romanian chords and tablatures, having been designed according to the principle: anyone who wants to can contribute, and everyone benefits. The archive contains songs with chords or tablatures from all styles and eras, ranging from mountain songs to the latest hits. According to our investigations, the site first appeared during 2006, offering a few hundred songs; the archive steadily expanded over the years to nowadays offer almost 10,000 songs. The website is not primarily intended as a repository of lyrics, but rather as an aid for musicians who play the guitar; but while guitar tabs are the main focus, we observed that most songs also featured the lyrics. Compared to other, larger websites such as *versuri.ro*, *tabulaturi.ro* offers quite rich metadata besides song lyrics, including the genre tags, the key, chords, guitar tablatures and piano transcripts. Because the website does not offer a public web API, we used web crawling techniques to obtain and parse the data. Alongside the lyrics and guitar tabs, the name of the song, the artist, and the genre were also obtained.

3.2. Labeling

What makes this site special lies both in its purpose and in the way it was built. First of all, the fact that it's a tablature site means it primarily contains songs that can be played on the guitar. While, in principle, any song can be played on the guitar—since it's made up of musical notes and can be rendered using chords—this creates a very visible bias in the representation of musical genres on the platform: folk dominates the site, with over 3000 entries, followed by rock and pop (including pop-rock, ska, and so on). But the fact that this is a tablature site—meant for people looking for musical styles compatible with chord-based accompaniment—means that from the beginning, the site included a genre category in its structure. In fact, it's the most relevant music-related crowdsourced database in Romania. While there are much larger similar sites like *versuri.ro*, these tend to be just for hosting lyrics, so genre tags probably seemed unnecessary to their creators and are therefore missing from the site. Of course, we would have preferred to work with the much larger archive on *versuri.ro*. But in the absence of genre tags, we used the archive on

tabulaturi.ro instead. In the future, the current work can be used to annotate genre-less lyrics in other datasets.

The archive also shares one more interesting feature. Beyond the fact that in music genres are quite changeable, through what is generally known as genre fusion, genre blending, or crossover & genre-busting (forms of multiple-genre composition, where songs often lie at the intersection of two or more genres, sharing key traits from each, like pop-rock or country-rock), our archive brought around invented genres by users or website. Beyond classical and historicized genres, we had ad hoc functional categories like “party music” or geographical ones like “music from the Republic of Moldavia.”

Another feature of the dataset is that each song could belong to multiple genres. Since we wished to avoid the complexity of working with a multi-label dataset and having to predict multiple genre labels per song, this situation was solved by adding a duplicate song to each genre label. This ensured that as much data was kept as possible, as eliminating these songs would have drastically reduced the dataset size, but at the same time arbitrarily assigning each song to only one genre would have been impossible to justify methodologically. The cost is added redundancy in the data, growing our dataset to a size of almost 14,000, an acceptable tradeoff for the current methodology.

The relatively uninformative nature of some genres (“Din Republica Moldova”—“From the Republic of Moldova” being, for instance, a geographical rather than stylistic description) and the high number of classes (42) along with a very imbalanced data distribution (see Figure 1) prompted us to merge some genres according to the mapping in Table 1. The new genre distribution after merging can be seen in Figure 2.

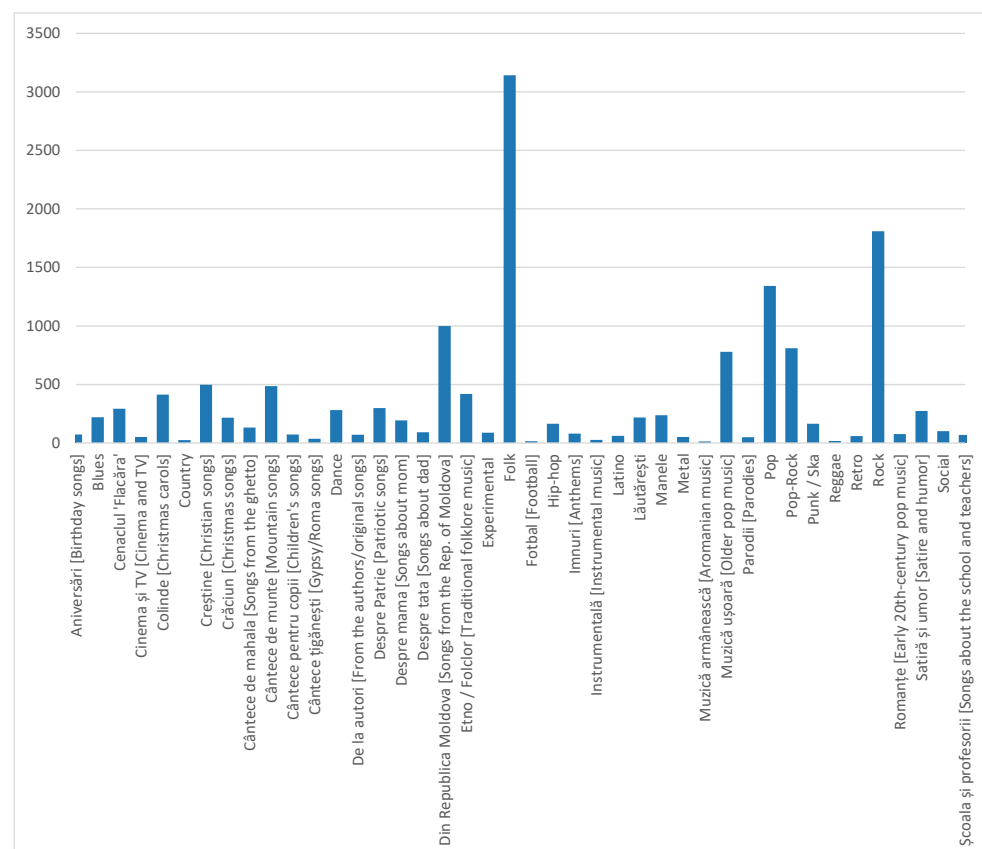


Figure 1. Initial genre distribution.

Table 1. Genre mappings.

Original Genre Label	Merged Genre Label	Notes on Context
Etno/Folclor	Muzica populara	Traditional Romanian folklore music
Aniversări	Altele	Birthday songs, sorted into Altele [<i>Other</i>]
Din Republica Moldova	Altele	From the Republic of Moldova
Muzică ușoară	Pop	Literally means “light music”, a type of pop music focused on romantic themes popular in the second half of the 20th century. Sometimes used in Romanian as a synonym for all popular music (used in opposition to <i>muzică cultă</i> , art music)
Despre mama	Altele	Songs about mom
Lăutărești	Manele/Lautareasca	<i>Lăutărească</i> is a type of traditional folk music played by professional Roma musicians; <i>Manele</i> is a form of modern pop music sung by Roma singers. The genres share some stylistic traits and some musicians have worked within both genres, thus the association
Instrumentală	Altele	Instrumental music
Pop-Rock	Pop	
Cântece pentru copii	Altele	Children’s songs
Dance	Pop	
Rock	Rock	
Romațe	Pop	An early form of pop music, these are romantic songs sung in a heightened emotional style, sometimes accompanied by piano and strings, popular in the early-to-mid 20th Century
Blues	Rock	
Școala și profesorii	Altele	Songs about school and teachers
Pop	Pop	
Folk	Folk	Folk (not to be confused with traditional folklore music) was inspired in its style from Western Folk music concentrating on the guitar and voice, and often tackling historical, literary or introspective themes

Table 1. Cont.

Original Genre Label	Merged Genre Label	Notes on Context
Country	Rock	
Social	Altele	
Crăciun	Muzica religioasa	Christmas songs and religious music, respectively
Cenaclul ‘Flacăra’	Folk	<i>Cenaclul ‘Flacăra’</i> was an influential youth culture circle during the last decades of the Communist period, mostly associated with the Folk genre
Satiră și umor	Altele	Humorous songs
Cântece de mahala	Manele/Lautareasca	Songs from marginalized communities/the ghetto
Cântece țigănești	Manele/Lautareasca	Gypsy (Roma) songs
Cântece de munte	Folk	Mountain songs
Parodii	Altele	Parodies
Imnuri	Altele	Anthems
Colinde	Muzica religioasa	Christmas Carols
Cinema și TV	Pop	Cinema and TV songs
Despre Patrie	Altele	About the Fatherland/patriotic songs
Muzică armânească	Muzica populara	Aromanian songs
Despre tata	Altele	Songs about dad
Creștine	Muzica religioasa	Christian songs
De la Autori	Altele	Original songs (provided by site users)
Experimental	Altele	
Manele	Manele/Lautareasca	
Hip-hop	Hip/Hop	
Latino	Pop	
Fotbal	Altele	Songs about football
Metal	Rock	
Punk/Ska	Rock	
Reggae	Rock	
Retro	Pop	

Note the presence and large proportion of the label “Altele” [*Others*]. Because of the heterogenous nature of this category, after some experimentation which proved its presence could be problematic for fitting unsophisticated classifiers such as logistic regression, we decided to exclude it from the experiments discussed in this paper. Also note that a prediction of “other” is not really informative, so it would have been quite misleading to count data classified as such towards the accuracy. Nevertheless, we kept it in the curated dataset for future use.

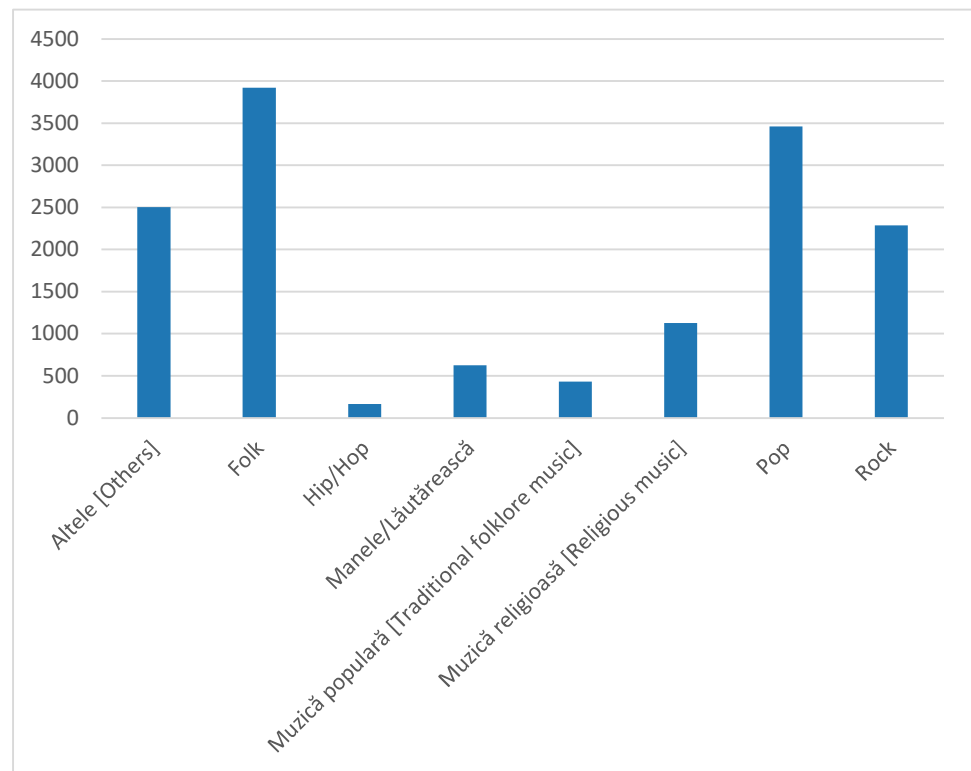


Figure 2. Genre distribution after merging.

3.3. Preprocessing

To ensure a uniform input data format, we applied cleaning and preprocessing for the lyrics of each song. For all texts, we removed all non-punctuation symbols, digits, and multiple whitespaces, and converted to lowercase. Additionally, since some lyrics contained guitar tablatures or chord symbols (see examples in Figure 3), we removed these as well. Since we are interested both in verse-level and phrase-level features, we preserved the verse structure by replacing newlines with a special token. For phrases, we split the text by the punctuation symbols “.”, “;”, “!”, “?”, “:”.

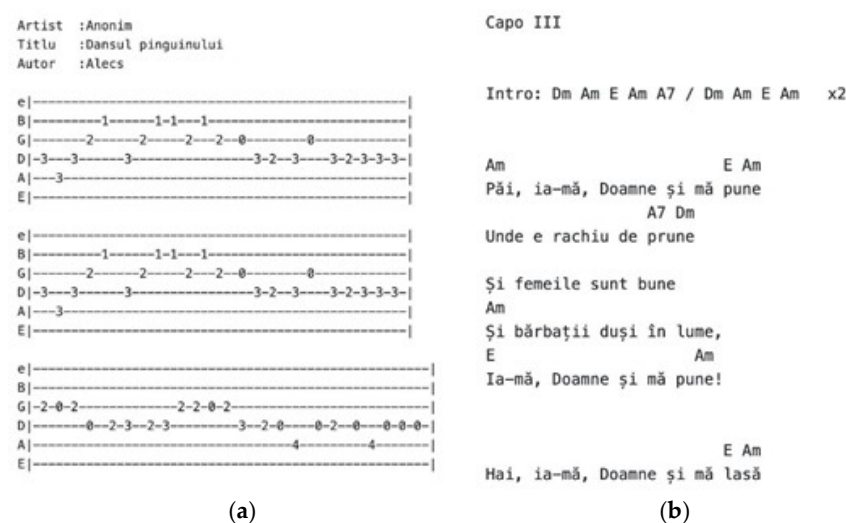


Figure 3. Examples of non-token artifacts (a) Tablatures inserted directly as text; (b) Lyrics overlaid with chords and using whitespace to suggest timing.

3.4. Lexical Analysis

Since some genres are associated in the popular imagination with ethnic minorities (Manele/Lăutărească, for example) [32,33] or with specific attitudes (Hip/Hop with misogynistic and homophobic attitudes), we sought to capture these lexical features of the text and give them special attention. For this purpose, we developed a lexical analysis library by compiling lists of three types of vulgarities: sexual, ethnic, and swear words. We obtained the most common vulgarities using Romanian-language Wiktionary (https://en.wiktionary.org/wiki/Category:Romanian_vulgarities (accessed on 5 December 2024)) as well as a survey of the Romania-focused subreddit r/Romania; out of these, for sexual slurs, we selected the ones with a gender- or sexual orientation-based negative connotation. For ethnic slurs, we used the list available on the Romanian-language Wiktionary (https://en.wiktionary.org/wiki/Category:Romanian_ethnic_slurs (accessed on 5 December 2024)). We then enriched our lexicon with common colloquial misspellings and variants of these, as well as the word's paradigm which we obtained manually using the popular Romanian dictionary website dexonline (<https://dexonline.ro/> (accessed on 18 December 2024)). The lexicon files are available in the dedicated repository (<https://github.com/erkovacs/2025-ro-song-lyrics-classification-public> (accessed on 5 November 2025)).

SentiLex is a popular multilingual library for sentiment analysis applications using a lexicon-based approach [14] which evaluates the sentiment of a text sample in two binary opposites: *positive* sentiment and *negative* sentiment. While it nominally supports Romanian as a language, during initial experimentation on our corpus it provided dissatisfying results and we sought to inspect it internally, ultimately deciding to rebalance the lexemes so as to increase its accuracy. This transformation began by modifying the word corpus on which the initial model was based on. Within this paper, we denote our modified version as SentiLex-v2. First, we expanded the dictionary by adding inflectional and paradigmatic forms. The reason for this is that the initial package requires that the text fed into it first be lemmatized. However, we have observed questionable performance from the most common lemmatization libraries such as nltk's SnowballStemmer for the Romanian language. Because the corpora are relatively small in size, manually adding these forms was feasible and we hoped for an increase in accuracy from this change. Then, we applied two types of direct transformations to the corpus: moving entries from one category to another (often because we decided that, at least in the context of artistic language, the new category was more appropriate); and elimination (either because the entries did not directly refer to any kind of sentiment or to any form of imaginary that could be directly associated with a sentiment, or because of problematic ambiguity). We believe that these issues are artifacts of the machine translation approach through which these corpora were obtained initially [14].

First, there was a whole set of words moved from positive to negative and, conversely, from negative to positive. We moved words like “ajunge” [*enough* or *to arrive*] from positive to negative, precisely because their meaning is closer to “that’s enough” or “that will do,” and they mostly appear in contexts expressing a kind of saturation toward a situation. Moreover, words like “permit” [*to allow*] or “permite” [*allows*] were moved because they define more of an arrogant or pessimistic attitude (“își permite” [*she/he has the nerve, has the audacity*]; “îmi permite” [*she/he allows me, puts me in a subordinate position where I need someone’s permission*]). “Acoperi” [*to cover*] was moved to negative because we could hardly think of a context in which it would have a positive connotation: although it either means “to lie for someone,” it most commonly refers to “covering” things with other things, metaphorically suggesting that the former are being forgotten or neglected. Similarly, “ascuțit” [*sharp*], even though it can define something positive, was moved to negative since

in most situations it carries the meaning of “irritating” or “dangerous” (a “voce ascuțită” [*sharp voice*] is an irritating voice, a “cuțit ascuțit” [*sharp knife*] is dangerous, etc.). As for the words “avocat” [*lawyer*] and “avocați” [*lawyers*], we moved them to negative precisely because their presence implies the existence of a non-amicable dispute—and thus a lack of agreement. “Bocet” [*wailing*] was moved because it means “weeping” and conveys a sense of sadness; it is not used in optimistic contexts. There were also a few ideological choices. We moved “economie” [*economy*] and “economic” [*economic*] to negative because they imply the existence of prior disorder that requires “calculation,” a rationalization of reality, whether social or emotional. “Fundas” [*defender*] was moved to negative because it denotes staying back, close to “codaș” [*laggard*]. Terms like “comoditate” [*comfort, ease*] or “comodități” [*comforts, conveniences*] were moved to negative because they describe a state of lack of initiative and action, often found in romantic reproaches, and resonate with the idea of “to take for granted.”

Secondly, words that carried a certain ambivalent emotional load were removed from SentiLex—meaning terms that could accompany and enhance both positive and negative sensations (attributes like “foarte” [*very*] or “mare” [*big*])—as well as words that, although they seem to refer to a factual state (“drept” [*right*], which can also mean “correct” or “just” [*fair*]), had a very inconclusive status due to their polysemous use. “Drept” [*right*] can appear in constructions like “e drept că” [*it is right that*], meaning “e adevărat că” [*it is true that*]. Because we are working with musical tracks, we removed words with positive connotations mainly in official contexts (such as legal language) and less in colloquial ones (e.g., “conform” [*in accordance with*]). Some words were considered plausible to function in both categories (and were therefore eliminated) because—even though they often build a positive imaginary—they are more commonly used in music as metaphors or referred to as lacking. The best example here is “afacere” [*business/deal*], which plays a role in expressing that a feeling is more transactional. This reveals a genre-based difference: “afacere” [*business/deal*] may be transactional in rock and dance music, but in manele it often carries a positive sentimental sense of ambition (see Valentino—“Cea mai bună afacere ești tu” [*You are the best deal*] from 2020) or praise of a life full of success (see Nicolae Guță—“Merg bine afacerile” [*Business is going well*] from 2006). Other terms, such as neologisms like “alimentație” [*nutrition*], rarely appear in lyrics and do not have a strong enough presence to be emotionally relevant as neologisms for either positive or negative sentiment. In selecting certain positive sentiments, some words were eliminated simply because they couldn’t be linguistically resolved by context (for example, “amant” [*lover*]). While the term can be found in songs that praise love, it can also describe a negative atmosphere in songs where the “amant” [*lover*] replaces and ruins a relationship. Similarly, we decided to remove “amendă” from both corpora, since it does not clearly convey either a literal meaning (a fine as punishment) or a metaphorical one (where the fine becomes a kind of erotic warning in songs).

Moving over 20 words from positive to negative and over 30 words from negative to positive, the elimination of over 200 words from both parts of the SentiLex corpus, alongside with the complete paradigm of the words themselves, helped us reach a 25% better sentiment analysis of the lyrics, as described under Section 4.2.

3.5. Feature Encoding

In computational approaches to poetry and song lyrics, structural and lexical features such as verse and phrase organization, vocabulary density, and sentiment play a crucial role in uncovering patterns that are often invisible to close reading alone. Metrics like verse length, phrase length, and stop word ratio allow us to capture rhythmic and prosodic tendencies, including phenomena such as enjambment, where the flow of meaning crosses

line boundaries—an important indicator of modern stylistic experimentation. At the lexical level, measures of word length, vocabulary density, and repetition provide insight into the balance between simplicity and complexity, as well as into the use of leitmotifs that rely on recurring words or expressions to establish identity within a genre. These features can be further nuanced by tracking the position of repetitions (e.g., at the beginning or end of lines), since placement often signals distinct rhetorical strategies. Finally, sentiment-based features—ranging from general polarity to the presence of profanity or racial slurs—open the way for analyzing affective registers and cultural positioning, showing how genres differ not only in form but also in the moods, attitudes, and social codes they project. Taken together, such features demonstrate how computational analysis can translate the traditional concerns of literary criticism—prosody, diction, rhetoric, and mood—into measurable variables, thereby enabling systematic comparisons across large corpora of poetic texts.

Because the scale of the features ranged widely, each feature was normalized using the well-known Equation (1), which replaces all components of the feature vector X with its z-score (its distance, measured in standard deviations, from the mean).

$$X_{norm} = \frac{X - \mu}{\sigma} \quad (1)$$

where X is a $1 \times n$ vector for a certain feature containing the value of that feature for all n observations, μ is the mean, and σ is the standard deviation.—denotes the elementwise subtraction operation (the right-hand operand is subtracted from all components of X), while the fraction line denotes elementwise division.

3.6. Models and Fitting

Initial classification was attempted using logistic regression. The reason for this choice was because we wanted an interpretable model which gives a clear meaning to the contribution of each feature and can quantify whether the feature meaningfully contributes to the prediction, since our aim was to identify which features are most relevant to which class. Initial poor fit suggested multicollinearity, that is, a degree of linear dependence within the features. Figure 4 shows the correlogram between the normalized features for the whole dataset. Certain features show high positive correlations of >0.7 (all_vulgarity_ratio was highly correlated with swear_word_ratio and ethnic_slur_ratio for instance); from Table 2 we observe this is somewhat expected since from the definitions of all_vulgarity_ratio, swear_word_ratio, sexual_slur_ratio, and ethnic_slur_ratio form close to a linearly dependent system. The following features also exhibit high correlations: char_count with repetitions_end and vocab_size; word_count with mean_phrase_length and mean_verse_length; and some of the sentiment-based features.

The elimination of all_vulgarity_ratio, repetitions_max_count, char_count, word_count, negative_sentiment and positive_sentiment eliminates significant (>0.5) absolute correlations between features, leading to better fit and a more parsimonious model which has no redundant features. Note that the black squares in the correlogram (Figure 4) only exhibit small negative correlations of around -0.25 , so they do not suggest other features need to be removed. vocab_size was deemed an important feature and kept despite showing some positive correlation with repetitions_end. The sentiment data obtained with the original version of SentiLex can also be removed safely because we already have the sentiment data obtained with SentiLex-v2, our modified version.

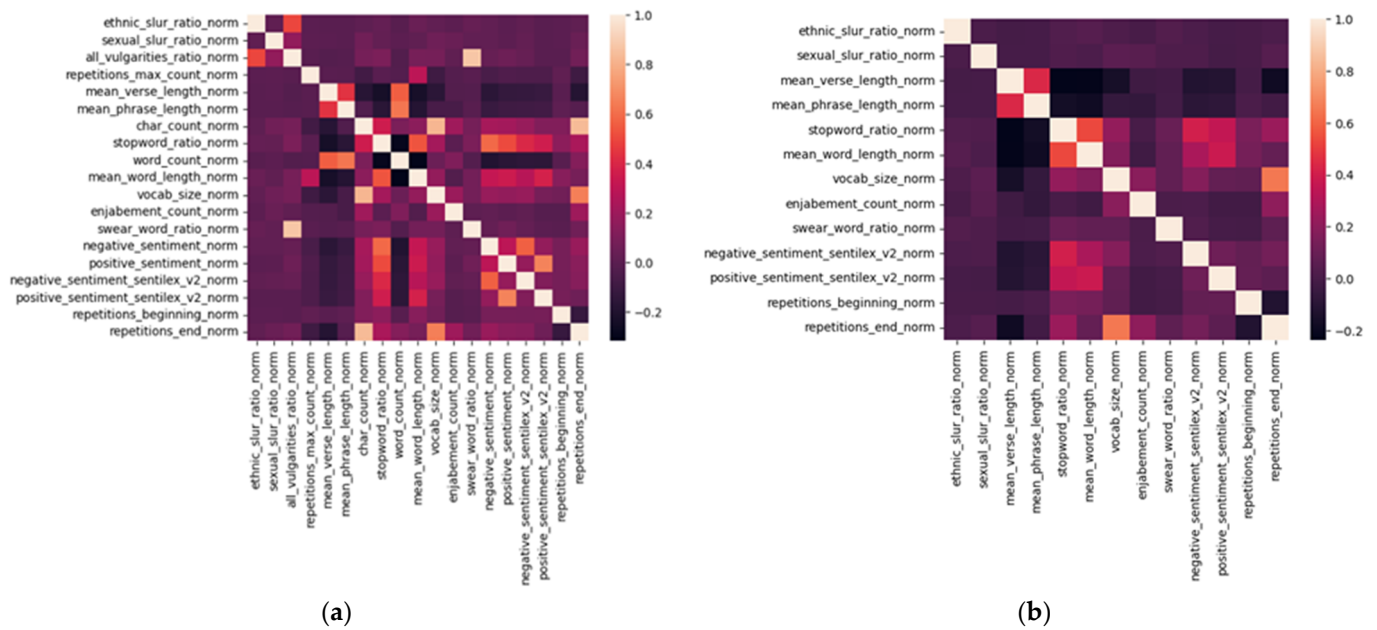


Figure 4. Correlations between features (a) Correlogram depicting all features (“norm” denotes the feature has been normalized); (b) Correlogram with highly correlated feature pairs removed.

Table 2. Features and description.

Feature	Definition
Lexical	
swear_word_ratio	number of swear words/number of words
ethnic_slur_ratio	number of ethnic slurs/number of words
sexual_slur_ratio	number of sexual slurs/number of words
all_vulgarities_ratio	(number of ethnic slurs + number of sexual slurs + number of swear words)/number of words
stopword_ratio	number of stopwords/number of words
word_count	number of words
mean_word_length	average length, in characters, of the words used in a song
vocab_size	total number of different words used in a song
Sentiment-based	
positive_sentiment	number of words with positive valence, based on SentiLex/number of words
negative_sentiment	number of words with negative valence, based on SentiLex/number of words
positive_sentiment_sentilex_v2	number of words with positive valence, based on SentiLex-v2/number of words
negative_sentiment_sentilex_v2	number of words with negative valence, based on SentiLex-v2/number of words
Stylistic and prosodic	
repetitions_max_count	maximum of the longest repeated sequence of the most frequent word out of each verse

Table 2. Cont.

Feature	Definition
repetitions_beginning	number of repetitions of the most frequent word in each verse considering only the first half of the verse
repetitions_end	number of repetitions of the most frequent word in each verse considering only the latter half of the verse
mean_verse_length	average verse length in words
mean_phrase_length	average phrase length in words
char_count	total number of characters in the song
enjabement_count	number of phrases that end/start within a verse instead of at the end of a verse

Since music genre classification is a multi-class classification problem, but logistic regression is a binary classification model, we used the ensemble method extension of logistic regression known as OVA, or One-Versus-All, in which k different models are fit for each of the k classes. For each class, the current class is marked with label 1, whereas all other classes will be labelled as 0. For inferencing, or classifying new data, each classifier will make a prediction; the final label will be chosen as the prediction with the highest probability out of the k predictions.

As can be seen in Figure 2, even after merging, the data is highly imbalanced, meaning that the proportions of the classes are highly unequal. To mitigate this issue, to which logistic regression is highly sensitive, we used a balanced undersampling approach for all classes as follows: half of the sample consisted of the positive class, with the other half consisting of a stratified random sample of all other classes (which for this particular setup we considered as the negative class, as explained above). This setup was replicated for all the classes. For estimating the accuracy of the models, we used a 80%–20% train–test split while setting the random seed on the system to a constant known value to ensure reproducibility. We used the statsmodels (<https://www.statsmodels.org/stable/index.html> (accessed on 13 January 2025)) python package for training the model.

4. Results

We present our results, in the form of a quantitative measurement for the improvements we have made to the sentiment analysis library SentiLex against a known baseline, and measure the accuracy of the logistic regression classifier in terms of the standard classification metrics. The most representative features are described and their contributions ranked for each class. Class profiles are built using radar plots of their feature representations.

4.1. SentiLex Improvements

We use the standard accuracy measures for all measurements. Accuracy (Equation (2)) captures how many of the predicted values were the same as the values in the data, but is highly susceptible to class imbalance. Precision (Equation (3)) captures how many of the predicted positive cases were indeed positive. Recall (Equation (4)) captures how many of the predicted cases were correct no matter if classified as positive or negative. Finally,

the F1 score (Equation (5)) balances out precision and recall giving a global measure of accuracy which is less susceptible to class imbalance.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

$$F_1\text{score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (5)$$

where TP is the number of true positives (predicted genre matches actual genre for current genre), TN is the number of true negatives (predicted negative for current genre and actual genre was another genre), FN is the number of false negatives (predicted negative for current genre but actual genre matches current genre), and FP is the number of false positives (predicted a genre but actual genre does not match).

To estimate the accuracy gain from our modifications to the SentiLex lexicon, we used the sentiment analysis dataset made available by Dumitrescu et al., consisting of product and film reviews [34].

We set up two parsers: one using the unmodified SentiLex lexicon files and the other using our new, modified files as described under Section 3.4. Otherwise, the pre-processing pipeline was identical: removal of all symbols, lowercasing, and stemming of the text using the standard Romanian Snowball Stemmer (<https://www.nltk.org/api/nltk.stem.SnowballStemmer.html?highlight=stopwords> (accessed on 7 January 2025)) available in the nltk python package. As noted above, this is no longer strictly necessary for SentiLex-v2, but it is still necessary for the original SentiLex. We then used the test portion of the benchmark data to estimate the accuracy of the two parsers. The net sentiment score for each text x_i was calculated as in Equation (6).

$$net\text{-}sentiment_{x_i} = positive\text{-}sentiment_{x_i} - negative\text{-}sentiment_{x_i} \quad (6)$$

Then, the class (0—negative, 1—positive) was assigned as follows in Equation (7):

$$sentiment\text{-}label_{x_i} = \begin{cases} 1, & net\text{-}sentiment_{x_i} > 0 \\ 0, & net\text{-}sentiment_{x_i} < 0 \\ -1, & net\text{-}sentiment_{x_i} = 0 \end{cases} \quad (7)$$

The data with a net-sentiment of 0, that is, those labelled with -1 , were removed so that only data with a nonzero net sentiment was used for the benchmark. We report the simple accuracy, precision and recall, and the F1 score for the two sets of lexicons in Table 3. The accuracy gain we see, an F1 score of 69.94% over the 55.81% baseline, represents a significant 25% increase, suggesting that our rebalancing measurably improved the lexicon.

Table 3. The benchmark results for SentiLex.

Lexicon	Accuracy	Precision	Recall	F1 Score
SentiLex [14]	61.48%	40.73%	88.63%	55.81%
SentiLex-v2	68.39%	61.59%	80.91%	69.94%

Note the large difference between the precision and the recall for both lexicons. This can be explained by the fact that any lexicon-based approach is limited in its ability to

recognize context. The words forming a sentence are not independent from each other, yet a lexicon-based system inherently treats them as such when calculating the average net sentiment; that means that the presence of certain words increases the sentiment score for that text sample no matter the context. Likewise, the presence of many negative words can counteract the presence of many positive ones. These issues account for the lower precision scores due to the generation of many false positives (positives in the sense of predicting a certain sentiment score, not as in predicting a *positive* sentiment) in texts where words with a certain sentiment score are used neutrally (as in describing them, for instance) or in an opposite sense (for example when being preceded by a negation which would invert the meaning, or when employing irony or sarcasm), with the actual meaning deducible only from context. The high recall values on the other hand show that the lexicon can identify most positive instances correctly, with a low proportion of false negatives. The increase in overall accuracy that we report is due to the significant improvement in precision for SentiLex-v2 (from 40.73% to 61.59%), suggesting that the contextual reasoning we used during the rebalancing does indeed generalize.

4.2. Genre Classification

Before turning to the detailed analysis of features, it is useful to pause on the accuracy of the classification model as well. Table 4 reports the values of accuracy, precision, recall, and F1 score for each genre as well as for the whole OVA ensemble.

Table 4. The accuracy on the test set for each class and for the ensemble as a whole.

Class	Accuracy	Precision	Recall	F1 Score
Muzica populara	71.84%	67.86%	85.39%	75.62%
Pop	59.32%	62.34%	53.70%	57.70%
Manele/Lautareasca	64.00%	69.92%	61.87%	65.65%
Rock	57.42%	58.35%	62.53%	60.37%
Folk	63.99%	63.74%	65.44%	64.58%
Muzica religioasa	68.14%	65.19%	81.97%	72.62%
Hip/Hop	54.54%	54.76%	67.65%	60.53%
Mean	62.75%	63.16%	68.37%	65.29%

The results point to several interesting tendencies. Genres such as popular and religious music obtain relatively high recall, meaning the model is effective at retrieving them, even if precision is more modest. By contrast, pop, rock, and hip-hop show significantly lower overall scores. For pop and rock this outcome is perhaps not surprising: these are highly general and hybridized categories, difficult to distinguish on the basis of lyrics alone. Hip-hop, however, presents a more puzzling case; despite being linguistically marked by density and complexity, it still achieves the weakest accuracy in our dataset. This suggests that lyrics by themselves may not provide sufficient dissimilarity in the feature space to allow robust prediction of genre (refer to Table 5 to assess the most predictive features for each genre).

The question, then, is not only why some genres resist accurate prediction, but also what can be learned from those with higher scores. If certain genres consistently emerge as classifiable based on textual features, they may indicate deeper structural consistencies that can serve as anchors for future models. Conversely, the low-scoring categories remind us of the inherent limits of lyrics-based classification and the need to complement

linguistic analysis with additional modalities such as instrumentation, performance style, or cultural context.

Table 5. The parameters of the model, shown together with the odds-ratio showing the associated feature's relative contribution towards classifying a sample as that genre. Only statistically significant features were included.

Model	Feature (Normalized)	Coefficient	Odds-Ratio
Muzica populara	mean_verse_length	1.1853	3.27
	stopword_ratio	0.3411	1.41
	mean_word_length	0.2005	1.22
	word_count	−0.8684	0.42
	vocab_size	−0.4737	0.62
	swear_word_ratio	0.1743	1.19
Pop	ethnic_slur_ratio	−0.0748	0.93
	stopword_ratio	0.1924	1.21
	mean_word_length	−0.2440	0.78
	vocab_size	−0.2921	0.75
	swear_word_ratio	0.0578	1.06
	negative_sentiment_sentilex_v2	−0.0864	0.92
	repetitions_beginning	0.1892	1.21
	repetitions_end	0.5047	1.66
	ethnic_slur_ratio	0.4597	1.58
	sexual_slur_ratio	7.5093	1824.94
Manele/Lautareasca	stopword_ratio	0.4955	1.64
	mean_word_length	−0.1770	0.84
	vocab_size	−0.8577	0.42
	swear_word_ratio	0.7599	2.14
	negative_sentiment_sentilex_v2	−0.2603	0.77
	repetitions_end	0.2055	1.23
	mean_phrase_length	0.1239	1.13
	stopword_ratio	−0.3424	0.71
	mean_word_length	0.2025	1.22
	vocab_size	−0.1091	0.90
Rock	swear_word_ratio	−0.1374	0.87
	positive_sentiment_sentilex_v2	−0.0904	0.91
	repetitions_beginning	0.0730	1.08
	mean_verse_length	−0.5696	0.57
	mean_phrase_length	−0.6225	0.54
	stopword_ratio	−0.1436	0.87
	vocab_size	0.6660	1.95
	swear_word_ratio	−0.0582	0.94
Folk			

Table 5. Cont.

Model	Feature (Normalized)	Coefficient	Odds-Ratio
Muzica religioasa	negative_sentiment_sentilex_v2	0.2485	1.28
	positive_sentiment_sentilex_v2	−0.1444	0.87
	repetitions_beginning	−0.2485	0.78
	repetitions_end	−0.4740	0.62
	mean_verse_length	−0.3007	0.74
	mean_phrase_length	0.2903	1.34
	stopword_ratio	−0.5145	0.60
	mean_word_length	0.4355	1.55
	vocab_size	−0.3784	0.68
	enjabement_count	−0.2578	0.77
Hip/Hop	negative_sentiment_sentilex_v2	−0.4002	0.67
	positive_sentiment_sentilex_v2	0.3175	1.37
	repetitions_end	−0.2445	0.78
	stopword_ratio	0.4305	1.54
	mean_word_length	−0.5938	0.55
	vocab_size	0.9149	2.50
	repetitions_beginning	−0.1945	0.82
	repetitions_end	−0.7356	0.48

What the chart (Figures 5 and 6) shows us is surprising: first of all, it shows us that musical genres really do differ clearly when it comes to lyrics. Although, in general, it is quite difficult to say whether a pop song is different from a rock song, the diagram shows that at least in some respects the two genres have opposite traits. We also looked at which traits are in general “in conflict” in order to consolidate some directions of observation on the diagrams. Precisely because they often seem “similar” (especially through the hybridization of genres and the mixing of pop artists with rock artists after the decline of metal), it is important to highlight some characteristics that can determine the genre.

Composure—syncopation versus serenity—may be measured through the frequency of linking words relative to phrase length and overall word count. Although this might appear to be a marginal feature of genre formation, it in fact reveals a deeper stylistic distinction. Rock and religious music emerge as the more serene genres: lyrical rather than epic, marked by fluidity and calm, with relatively few stop words interrupting the flow of discourse. By contrast, hip-hop, pop, manele, and folklore cluster on the syncopated side of the spectrum: more narrative and unsettled, characterized by fragmentation, abruptness, and a heavy reliance on filler and connective words. Pop and folk, situated between these two poles, suggest the ambiguity of hybrid forms—at once epic and broken, yet still tethered to a degree of lyrical composure.

Density—complex versus simple—may be understood as the ratio between vocabulary size and word length. Our data shows that the only genres in which vocabulary appears relatively poorer, when measured against word length and total word count, are rock (though the difference is minor), folklore, manele, and religious music. At one end of the spectrum, the more complex genres—hip-hop and folk, with pop positioned closer to them—display richer vocabularies in proportion to the words they employ. At the

other end, the simpler genres—religious, manele, folklore, and, to some extent, rock—rely on more limited vocabularies relative to word length. This finding invites an important observation: rock music, particularly in the former communist states, acquired a kind of noble aura, celebrated as a symbol of the West and of cultural freedom, often staking its claim to superiority in open and not always cordial opposition to manele and folklore. Yet, when measured by this most “elitist” of criteria—the density of vocabulary—rock proves to be aligned more closely with religious, manele, and folklore traditions than with pop or hip-hop.

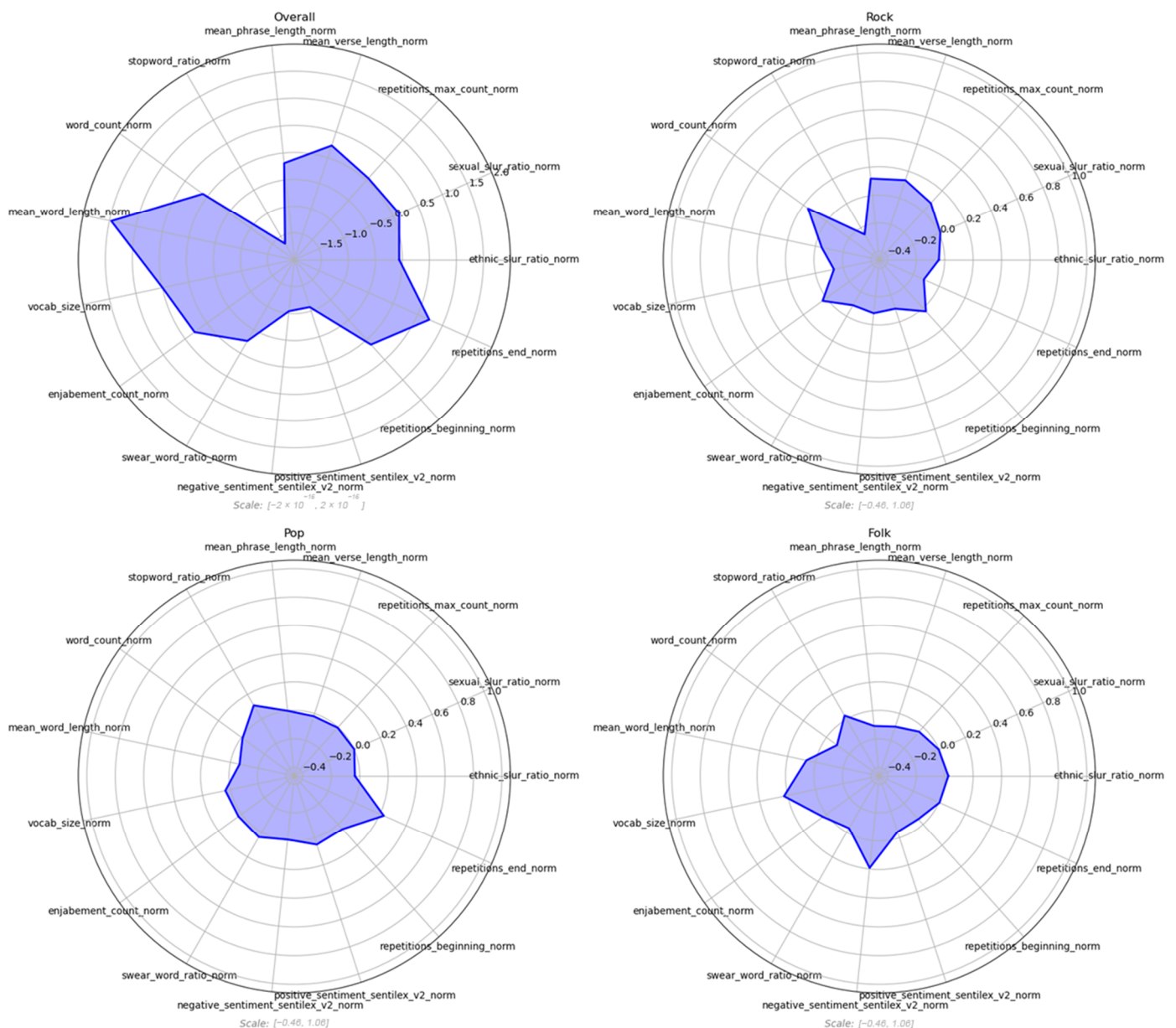


Figure 5. Feature profiles shown for the entire dataset and rock, pop and folk, respectively. The profiles were obtained by averaging features across all songs in the genre and showing the values on a common scale using polar coordinates.

Ritualism—*anaphoric* versus *cataphoric*—distinguishes musical genres through the specificity of their repetitive structures. As the diagram illustrates, certain genres tend to reiterate the beginnings of verses, a pattern we describe as anaphoric ritualism and associate with repetition_beginning (rock, manele, folklore, and religious music). Others emphasize

the endings of verses, exhibiting cataphoric ritualism, most clearly represented by pop and its preference for emphatic closure through repeated refrains. Still others maintain a relative equilibrium between initial and terminal repetition, producing an ambivalent ritualism characteristic of folk and hip-hop. What is striking, once more, is that rock, often positioned culturally in opposition to indigenous and popular traditions, nevertheless appears typologically aligned with folklore, manele, and religious music in its reliance on anaphoric repetition.

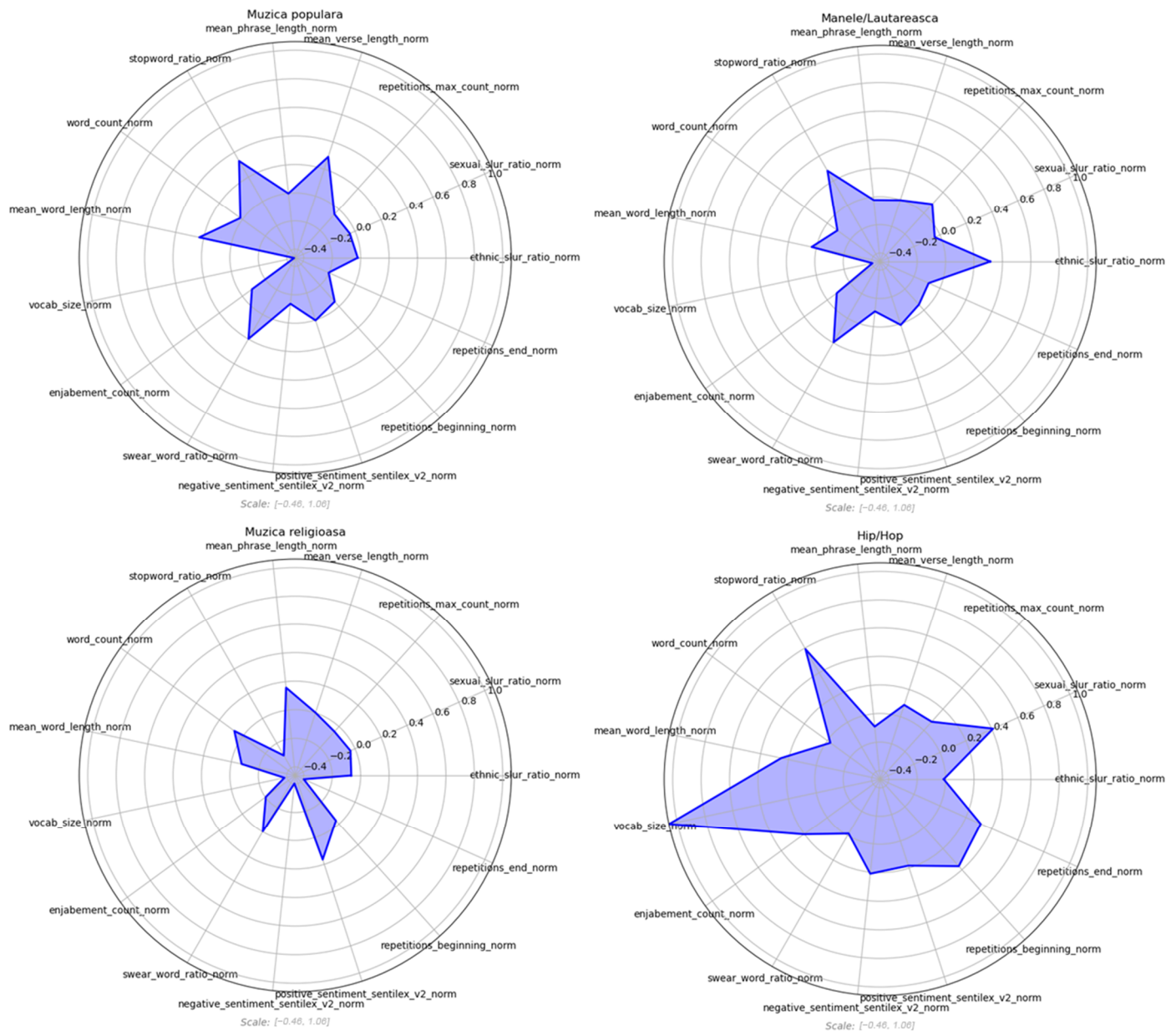


Figure 6. Feature profiles (continued). Figures are shown for traditional popular music, manele/lautareasca, religious music, and hip-hop.

Sentimentalism—ambivalent versus pathetic—can be traced through the ratio of positive to negative sentiments. Although in many cases the balance between positive and negative emotions is relatively even (with rock and pop tending slightly toward positivity, and hip-hop leaning slightly toward negativity), sentiment analysis nevertheless yields several revealing distinctions. Positive affect appears more consistently in genres rooted in indigenous traditions—popular, manele, and religious music—while negativity is more

strongly concentrated in a single imported genre, namely folk. Thus, even if the differences remain modest, the pattern is clear: the more ambivalent genres, balancing between positive and negative registers, include manele, pop, and folklore, while rock and hip-hop occupy the negative side of the spectrum. By contrast, the more pathetic genres—those which heightened sentiment in one dominant direction—are religious and folklore on the positive side, and folk on the negative. Once again, it is striking to observe rock positioned in close proximity to traditions with which it has often been considered culturally incompatible.

In terms of vocabulary, things are interesting as well. What we know for sure is that the only musical genres in which vocabulary is poorer relative to word length and number of words are rock (even if the difference is small), folklore, manele, and religious music. Here it is worth making the small observation that rock acquired, especially in the former communist states, a noble status, as it recalled the West and the “free” and “diverse” culture of Western Europe; thus, it always boasted of being superior to the others, often entering into public, open, and not at all pleasant conflict with popular and *lăutărească*. What our study shows, however, is that from the most elitist point of view—the complexity of vocabulary in relation to the number and length of words—rock is closer to religious and *lăutărească* music than to folk or hip-hop. Moreover, this is the second time that rock shares a typological trait with religious music. The most complex musical genre here is hip-hop, by far, earning its title as the genre of complexity *par excellence*. It is followed by folk (precisely because most songs here are actually musical transpositions of poems from high literary culture) and pop.

5. Discussion

We interpret the abovementioned results, revealing counter-intuitive similarities between traditionally “opposed” genres. We identify theoretical categories based on the features of each genre to explain their similarities and differences. Then, we note the limitations of the current study and outline potential areas for future research.

5.1. Analysis

If we try to describe how exactly the three main musical genres in our list can be recognized (folk with almost 4000 entries, pop with almost 3500, rock with almost 2500), it is absolutely necessary to look at what fundamentally distinguishes each one from the other two. What the chart (Figures 5 and 6) makes immediately apparent is that musical genres, though often perceived as overlapping or hybrid, in fact diverge quite distinctly at the level of lyrical structure. While the decline of metal and the subsequent cross-pollination between pop and rock artists has blurred genre boundaries in Romania, the analysis of composure, density, ritualism, and sentimentalism allows us to uncover consistent patterns of differentiation. These patterns are particularly striking when considered against cultural self-perceptions of the genres.

Rock provides the most telling example. Despite its historical role as a symbol of Western modernity and cultural emancipation in the former communist states—often asserted in deliberate opposition to manele, folklore, and religious traditions—our analysis reveals that, lyrically, rock aligns typologically with precisely these genres it has long repudiated. In terms of composure, rock belongs to the serene pole, marked by fluidity, calm, and an absence of connective clutter. In density, its vocabulary is simpler, closer to religious and folkloric registers than to the complexity of hip-hop or the literary stylization of folk. Its ritualism is predominantly anaphoric, privileging repetition at the beginning of lines, once again situating it alongside manele, folklore, and religious music. Finally, its sentimentalism shows a subtle leaning toward negativity, placing it closer to hip-hop than to pop, though still ambivalent. Taken together, these features suggest that rock,

while proclaiming its difference, imitates the linguistic architecture of traditions it sought to distinguish itself from. It is therefore not surprising that, to date, there are no real fusion experiments between rock and manele, folklore, or religious music: the proximity is structural rather than intentional.

Pop, by contrast, defines itself almost entirely in opposition to rock. With the exception of sentimentalism—where both remain broadly neutral—pop systematically occupies the opposite pole: syncopated rather than serene, complex rather than simple, cataphoric rather than anaphoric. Its language relies on fragmentation and emphatic repetition of endings, distancing it from the solemn, narrative qualities of rock and its culturally “older” allies. Together with hip-hop, pop thus represents a younger lyrical formation in Romanian culture, one that sought emancipation not by borrowing from rock but by constructing a distinct vocabulary of jitters, complexity, cataphoric ritualism, and sentimental ambivalence. This oppositional stance underscores why, despite surface-level hybridization, the two genres rarely overlap in deeper structural terms.

Folk occupies yet another position, one rooted in its genesis as a musical adaptation of literary poetry. Its lyrical identity is therefore shaped by jitters and complexity, reflective of its literary origin, while its ritualism remains ambivalent, balancing repetitions at both beginnings and endings. In sentimental terms, folk gravitates toward the pathetic, heightening affect in one direction—most often negativity. This combination has endowed folk with a reputation as the most “noble” genre, particularly for a generation of listeners shaped by the socialist aestheticism of the Ceaușescu era, where literature remained a privileged, if rigid, cultural domain. Folk thus appears both as a product of high-cultural transposition and as an ambivalent player within the wider field of Romanian musical genres.

In sum, the comparative analysis of composure, density, ritualism, and sentimentalism highlights not only the internal logic of each genre but also the paradoxes of their cultural positioning. Rock, the ostensible herald of Western modernity, turns out to be structurally aligned with the very traditions it disavows. Pop, in contrast, consolidates its identity in direct opposition to rock, defining itself through linguistic and structural innovation. Folk, meanwhile, inherits the authority of literature, yet situates itself ambivalently between complexity and affective intensity. Together, these findings complicate any simple mapping of genre by cultural affiliation, showing instead that lyrical structures may reveal unexpected proximities and oppositions, often at odds with the self-understanding of the genres themselves. Of course, the model we propose recognizes musical genres precisely by combining all the measurable variables we described. In fact, far from speculating around a single variable, the model proposes the analysis of a complex dataset through which it can decide on the highest probability that a song belongs to a certain genre.

5.2. Limitations and Future Work

Music genres have a hierarchical structure, in the sense that more general genres such as classical music or rock can contain sub-genres such as baroque or metal, and so on. Li and Ogihara propose a taxonomy of genres instead of a flat classification, applying it to the problem of content-based music classification [35]. Due to the nature and provenance of our data, however, our genres lack these hierarchical relationships. In the future, we aim to improve our work by including a taxonomical relationship structure in our genre labelling process.

Furthermore, the improvements to SentiLex were validated only on a product and movie review dataset, not on lyrical data. This means its suitability for sentiment analysis in lyrics has yet to be assessed. In the future we intend to perform manual annotation on lyric data for validation to ensure the improved lexicon’s effectiveness in a musical context.

To build an easily interpretable model, we have used logistic regression classification. The accuracy of this classical model is not comparable to the state of the art. Moreover, the presence of identical feature-space representations in multiple classes, an artifact of how we dealt with the multiple genre labels in the source data, is a known intrinsic weakness of logistic regression. The theoretical best probability it can assign to these examples is equal probability for each class they are part of, which is not meaningful, further degrading overall accuracy. In our future work, using deep learning models, we will be able to achieve higher accuracy than we have demonstrated, as these models are well-suited to fuzzy decision boundaries. We intend to fine-tune Transformers-based models such as BERT and RoBERTa on our dataset to achieve higher accuracy [26]. We also wish to explore multi-label classification approaches using these more advanced models to overcome the abovementioned challenges. The models we trained during the experiments described in this paper will furthermore be used to automatically label a larger lyrics dataset which we are in the process of gathering. This will help in the discovery of different lyrical themes within Romanian music genres, helping us to map out the local musical landscape using a statistical approach.

The categories we introduced—composure, density, ritualism, and sentimentalism—open avenues for further research that go beyond the immediate task of genre identification. Each category can be refined into a broader axis of analysis that speaks not only to genre classification but also to cultural history and aesthetics. Composure, for instance, could be examined diachronically, to see how the balance between narrative disruption and lyrical serenity shifts across decades or in response to sociopolitical change. Density invites comparative work between literary and musical corpora, testing whether genres that appropriate poetry or high literature consistently display more elaborate vocabularies, or whether the complexity of hip-hop constitutes a unique case of vernacular sophistication. Ritualism, with its distinction between anaphoric and cataphoric structures, raises the possibility of mapping genres through their rhetorical strategies, linking song lyrics to traditions of oral poetry, prayer, or propaganda. Finally, sentimentalism suggests a path for integrating computational analysis with cultural studies, as the distribution of positive and negative sentiments across genres may reflect deeper ideological divisions between “indigenous” and “imported” traditions. Taken together, these categories provide more than a technical framework for machine classification: they outline a conceptual toolkit for understanding how genres build identity through language. Future work will need to test their robustness across other corpora, explore their explanatory power for hybrid or fusion genres, and consider whether new dimensions are required to capture features that elude our current model. What becomes clear is that genre classification cannot be treated as a static taxonomy but rather as an evolving field in which linguistic evidence, cultural positioning, and historical context intersect in unpredictable ways.

6. Conclusions

We set out to answer the question: can we tell the song’s genre based on its lyrics using prosodic, stylistic, syntactic, and sentiment-based features? To find out, we gathered approximately 14,000 songs from the website *tabulaturi.ro*, and we built improved tools for Romanian language natural language processing in order to perform lexical analysis. Our analysis tool was designed to be well suited to song lyrics or poetry, being able to encode a set of 17 linguistic features. In addition, we built lexical analysis tools for profanity-based features and improved the SentiLex sentiment analysis library by modifying its lexicon to overcome its current limitations. We used a logistic regression classifier to distinguish between genres and built feature profiles for each of the genres, noting significant differences between the average values of their features.

The experimental results suggest that it is indeed possible to predict the song's genre from its lyrics using the prosodic, stylistic, syntactic, and sentiment-based features we propose, but the relationship between lyrical structure and genre is not straightforward. The results also clearly show that much of the common wisdom stereotyping certain genres as having certain characteristics or being overly concerned with a specific subject matter finds little support in the data. Rather, second-order characteristics of the genres, such as what we have termed composure, density, ritualism and sentimentalism, emerge across genres in a complex way, sometimes bringing historically, musically and sociologically opposed genres close together while separating genres often seen as going naturally together.

What emerged from this process is more than a typology of Romanian musical genres. Our results reveal that while lyrics can indeed be used to identify genre with measurable accuracy, they also complicate genre boundaries in unexpected ways. Genres that define themselves in cultural opposition often display striking structural similarities, and genres that claim literary or aesthetic nobility may, in fact, cluster linguistically with those seen as “popular” or “low.” These findings point to a central difficulty for the future of genre classification: lyrics are not merely markers of affiliation but also sites of hybridity, imitation, and paradox. The challenge going forward is not only to refine computational tools to detect such nuances but also to rethink what “genre” means when linguistic evidence undermines the very cultural distinctions it was supposed to uphold.

Author Contributions: Conceptualization, E.-R.K. and S.B.; methodology, E.-R.K.; software, E.-R.K.; validation, E.-R.K. and S.B.; formal analysis, S.B.; investigation, E.-R.K. and S.B.; resources, S.B.; data curation, E.-R.K. and S.B.; writing—original draft preparation, E.-R.K.; writing—review and editing, E.-R.K. and S.B.; visualization, E.-R.K.; supervision, S.B.; project administration, E.-R.K. and S.B.; funding acquisition, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No 101001710.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at <https://github.com/erkovacs/2025-ro-song-lyrics-classification-public> (accessed on 19 October 2025).

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

References

1. Fabbri, F. A Theory of Musical Genres: Two Applications. In Proceedings of the Popular Music Perspectives, First International Conference on Popular Music Studies, Göteborg and Exeter: International Association for the Study of Popular Music, Amsterdam, The Netherlands, 22–26 June 1981; Available online: <https://www.semanticscholar.org/paper/1-A-THEORY-OF-MUSICAL-GENRES-%3A-TWO-APPLICATIONS-Fabbri/feb8161666c893ed53a22eec9a3e7bba1f54fd57> (accessed on 14 November 2025).
2. Holt, F. *Genre in Popular Music*; The University of Chicago Press: Chicago, IL, USA, 2007; ISBN 0-226-35039-8.
3. Bronstein, M.; Droge, A.; Fredner, E.; Heuser, R.; Manshel, X.; Nomura, N.; Porter, J.D.; Walser, H. Microgenres. Available online: <https://litlab.stanford.edu/projects/microgenres/> (accessed on 3 October 2025).
4. Udrea, A.C.; Ruseti, S.; Pojoga, V.; Baghiu, S.; Terian, A.; Dascalu, M. Identifying Literary Microgenres and Writing Style Differences in Romanian Novels with ReaderBench and Large Language Models. *Future Internet* **2025**, *17*, 397. [CrossRef]
5. Poell, T.; Nieborg, D.; Duffy, B.E.; Prey, R.; Cunningham, S. The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity. *New Media Soc.* **2018**, *20*, 4275–4292. [CrossRef]
6. Washington, C.J.I. “It’s All the Same”: Genre Generalization in the American Music Industry. Master’s Thesis, Florida State University, Tallahassee, FL, USA, 2024.
7. Blakeley, R. Against the Stream: Niche Music Streaming Services and the Streaming Paradigm. Ph.D. Thesis, University of Rochester, New York, NY, USA, 2024.

8. Bowsher, A. Authenticity and the Commodity: Physical Music Media and the Independent Music Marketplace. Ph.D. Thesis, University of Oxford, Oxford, UK, 2014.
9. Baghiu, Ș. Apartenența multiplă de subgen: O propunere pentru istoria formelor romanesti. *Transilvania* **2022**, *11–12*, 45–49. [CrossRef]
10. Griffiths, D. From Lyric to Anti-Lyric: Analyzing the Words in Pop Song. In *Analyzing Popular Music*; Cambridge University Press: Cambridge, UK, 2003.
11. Bories, A.-S.; Plecháč, P.; Fabo, P.R. *Computational Stylistics in Poetry, Prose, and Drama*; De Gruyter: Berlin, Germany, 2022; ISBN 978-3-11-078150-2.
12. Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. Computational Social Science. *Science* **2009**, *323*, 721–723. [CrossRef] [PubMed]
13. Kovacs, E.-R.; Cofas, L.-A.; Delcea, C. A Deep Learning Approach to Fine-Grained Political Ideology Classification on Social Media Texts. In *Computational Collective Intelligence*; Nguyen, N.T., Franczyk, B., Ludwig, A., Núñez, M., Treur, J., Vossen, G., Kozierekiewicz, A., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 3–14.
14. Chen, Y.; Skiena, S. Building Sentiment Lexicons for All Major Languages. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Toutanova, K., Wu, H., Eds.; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 383–389.
15. Li, T.; Ogiwara, M.; Li, Q. A Comparative Study on Content-Based Music Genre Classification. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, ON, Canada, 28 July–1 August 2003; Association for Computing Machinery: New York, NY, USA, 2003; pp. 282–289.
16. Aguiar, L.; Martens, B. Digital Music Consumption on the Internet: Evidence from Clickstream Data. *Inf. Econ. Policy* **2016**, *34*, 27–43. [CrossRef]
17. Jones, S. Music and the Internet. In *The Handbook of Internet Studies*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2011; pp. 440–451. ISBN 978-1-4443-1486-1.
18. McKay, C.; Fujinaga, I. Improving Automatic Music Classification Performance by Extracting Features from Different Types of Data. In Proceedings of the International Conference on Multimedia Information Retrieval, Philadelphia, PA, USA, 29–31 March 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 257–266.
19. Sijbesma, D. Evaluating Audio Feature Importances and Machine Learning Models to Enhance Music Genre Classification and Recommendations. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands, 2024.
20. Flederius, D. Enhancing Music Genre Classification with Neural Networks by Using Extracted Musical Features. Available online: <https://purl.utwente.nl/essays/80549> (accessed on 14 November 2025).
21. Pelchat, N.; Gelowitz, C.M. Neural Network Music Genre Classification. *Can. J. Electr. Comput. Eng.* **2020**, *43*, 170–173. [CrossRef]
22. Tsaptsinos, A. Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network. *arXiv* **2017**, arXiv:1707.04678. [CrossRef]
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473. [CrossRef]
24. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; Knight, K., Nenkova, A., Rambow, O., Eds.; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 1480–1489.
25. Singhi, A.; Brown, D.G. On Cultural, Textual and Experiential Aspects of Music Mood. In Proceedings of the International Society for Music Information Retrieval Conference, Taipei, Taiwan, 27–31 October 2014. Available online: <https://zenodo.org/records/1417391> (accessed on 14 November 2025).
26. Akalp, H.; Furkan Cigdem, E.; Yilmaz, S.; Bolucu, N.; Can, B. Language Representation Models for Music Genre Classification Using Lyrics. In Proceedings of the 2021 International Symposium on Electrical, Electronics and Information Engineering, Seoul, Republic of Korea, 19–21 February 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 408–414.
27. Mayer, R.; Neumayer, R.; Rauber, A. Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In Proceedings of the ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008. Available online: https://www.ifs.tuwien.ac.at/mayer/publications/pdf/may_ismir08.pdf (accessed on 14 November 2025).
28. Mayer, R.; Rauber, A. Music Genre Classification by Ensembles of Audio and Lyrics Features. In Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, FL, USA, 24–28 October 2011. Available online: <https://archives.ismir.net/ismir2011/paper/000127.pdf> (accessed on 14 November 2025).
29. Li, Y.; Zhang, Z.; Ding, H.; Chang, L. Music Genre Classification Based on Fusing Audio and Lyric Information. *Multimed. Tools Appl.* **2023**, *82*, 20157–20176. [CrossRef]

30. Watanabe, K.; Goto, M. Lyrics Information Processing: Analysis, Generation, and Applications. In Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), Online, 16 October 2020; Oramas, S., Espinosa-Anke, L., Epure, E., Jones, R., Sordo, M., Quadrana, M., Watanabe, K., Eds.; Association for Computational Linguistics: Vienna, Austria, 2020; pp. 6–12.
31. Moretti, F. *Distant Reading*; Verso Books: London, UK; New York, NY, USA, 2013; ISBN 1-78168-084-1.
32. Schiop, A. *Smecherie Și Lume Rea. Universul Social al Manelelor*; Cartier: Bucharest, Romania, 2017; ISBN 978-9975-86-032-1.
33. Beissinger, M.; Rădulescu, S.; Giurchescu, A. *Manele in Romania: Cultural Expression and Social Meaning in Balkan Popular Music*; Rowman & Littlefield Publishers: Lanham, MD, USA, 2016.
34. Dumitrescu, S.; Avram, A.-M.; Pyysalo, S. The Birth of Romanian BERT. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Punta Cana, Dominican Republic, 16–20 November 2020; Association for Computational Linguistics: Vienna, Austria, 2020; pp. 4324–4328.
35. Li, T.; Ogihara, M. Music Genre Classification with Taxonomy. In Proceedings of the ICASSP '05. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 23–23 March 2005; Volume 5, pp. v/197–v/200.

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.