JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710-1606
Volume 07 Issue 01
2024

Research Article
https://doi.org/10.56979/701/2024

Analysis and Clustering of Pakistani Music by Lyrics: A Study of CokeStudio Pakistan

Zia Ur Rahman[1], Muhammad Imran Khan Khalil[1*], Asif Nawaz[2], Izaz Ahmad Khan[3], Naveed Jan[4], and Sheeraz Ahmad[5]

[1]University of Engineering & Technology, Peshawar, 25000, Pakistan.
[2]Faculty of Electrical Engineering, Engineering Technology and Sciences Division, Higher Colleges of Technology, United Arab Emirates, 16062.
[3]Bacha Khan University, Charsadda, 24420, Pakistan.
[4]Shuhada-e-APS, University of Technology, Nowshera, Pakistan.
[5]Iqra National University, Peshawar, 25000, Pakistan.
[*]Corresponding Author: Muhammad Imran Khan Khalil. Email: imrankhalil@uetpeshawar.edu.pk

Received: February 21, 2024 **Accepted:** May 19, 2024 **Published:** June 01, 2024

Abstract: This research explores the application of unsupervised learning techniques to categorize and understand the lyrical content of CokeStudio songs. In a world where music transcends cultural boundaries, this study delves into the rich linguistic tapestry of lyrics, unraveling emotions, themes, and cultural nuances. We begin by employing Natural Language Processing (NLP) and analysis techniques to uncover the emotional underpinnings of these lyrical compositions. This emotional layering becomes the foundation for the subsequent clustering process. Multiple unsupervised learning algorithms, including K-Means, Hierarchical Clustering, and DBSCAN, are employed to categorize songs into thematic clusters. The quality of these clusters is assessed using the silhouette score, with the optimal number of clusters determined as 5, achieving a score of 0.41641. Furthermore, we develop a robust classification model utilizing machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Multinomial Naive Bayes for evaluation of our clustering. This model assigns CokeStudio songs to thematic clusters based on the results of topic modeling, enhancing our understanding of the cultural and emotional dimensions of these compositions. Logistic Regression, with SMOTE applied to NMF values, emerges as the best-performing model, achieving an impressive testing score of 89.47%. The research findings not only illuminate the intricate emotions and narratives woven into CokeStudio songs but also emphasize the practical application of machine learning in music analysis. By identifying and classifying thematic clusters within song lyrics, this study enriches our comprehension of cultural expressions through music and opens avenues for personalized music recommendations.

Keywords: Clustering; Lyrics Analysis; Cultural Exploration; Unsupervised Learning; Text Classification; Processing; Music, Natural Language; CokeStudio.

1. Introduction

This Music, a universal language, transcends geographical, linguistic, and cultural boundaries. It has the power to evoke emotions, tell stories, and reflect the rich tapestry of human experiences. Within the realm of music, CokeStudio stands as an epitome of artistic collaboration and cultural fusion, bringing together diverse musical traditions from South Asia. CokeStudio, an enduring platform that bridges the gap between traditional and contemporary sounds by fusing musical cultures and genres, is one of the astonishing creations of this modern music scene. According to Abid et al. [1], CokeStudio has significantly transformed and revitalised the Pakistani music scene while providing a vital platform for well-known artists.

This research embarks on a journey into the heart of CokeStudio's melodies, seeking to uncover the hidden treasures within its lyrical compositions. As music enthusiasts, researchers, and data scientists, we are drawn to explore the profound emotions, themes, and cultural nuances embedded in these songs' lyrics. Text mining and text categorization were utilised by Yang and Lee [2] to determine the sentiment of music. In a world increasingly defined by data-driven insights, the fusion of music and data science presents an exciting opportunity.

The primary aim of this study is to leverage state-of-the-art Natural Language Processing (NLP) techniques and unsupervised learning algorithms for the analysis and categorization of CokeStudio songs based on their textual content. By doing so, we intend to achieve several key objectives:

- Thematic Exploration: We endeavor to unveil the underlying themes and emotions woven into the lyrical fabric of CokeStudio songs. Each song is a narrative in itself, expressing sentiments, stories, and cultural intricacies.
- Clustering and Classification: Through advanced clustering algorithms, we aim to group songs into thematic clusters, allowing us to discern patterns and commonalities in the lyrical content. Furthermore, we have develop a robust classification model that categorizes songs into these clusters, enhancing our understanding of the cultural and emotional dimensions of these compositions.
- Practical Application: Beyond the realm of music analysis, this research holds the potential for practical applications such as personalized music recommendations, enhancing user experiences in music streaming platforms.

The journey we embark upon is a testament to the evolving synergy between art and technology. It is an exploration of the hidden layers beneath the surface of music, where data-driven insights harmonize with artistic expression. As we delve into the melodies and lyrics of CokeStudio, we aim to not only enrich our understanding of music but also to contribute to the broader fields of Natural Language Processing and machine learning.

In this pursuit, we combine the soulful rhythms of CokeStudio with the precision of data science, ushering in a new era of music analysis and appreciation. The flow of the work done as show in the below figure 1.

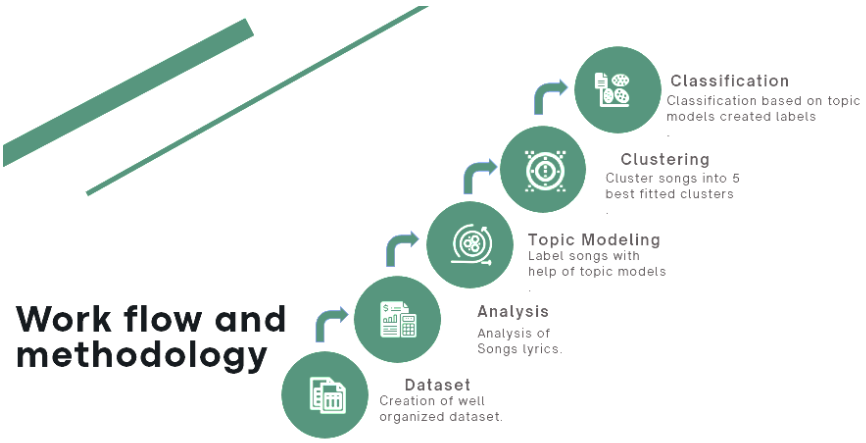


Figure 1. Flow of work done

2. Literature Review

Music is an expressive medium that cuts beyond linguistic, cultural, and geographical barriers. Applying unsupervised machine learning clustering algorithms to analyze and comprehend many elements of music, such as genre categorization, emotion detection, and theme grouping of song lyrics, has gained popularity in recent years. Without the need for labelled training data, unsupervised machine learning approaches seek to identify patterns and structures within huge datasets. The application of unsupervised machine learning clustering for music analysis, particularly in relation to song lyrics, is the main topic of this overview of the literature [4] [34].

Studies like those by De Masi [3] and Ali & Siddiqui [4] show how useful methods like Support Vector Machine classifiers and tf-idf features are for reaching high accuracy in genre categorization. Bergelid [5] tackles the problem of genre classification for explicit music material and discovers that SVM with TF-IDF

vectorization works better than alternative setups. In their study on the toxicity classification of song lyrics, Siddique et al. [6] discovered that Random Forest was incredibly successful.

Emotion detection from song lyrics is explored by Agrawal and An [7], Malheiro [8], and Nandwani and Verma [9], each proposing different methods like context-based approaches and sentiment analysis. Thematic grouping of songs based on lyrics is addressed by Rarasati [10], who achieved high accuracy in classifying themes like love and friendship using K-Means clustering[29].

Unsupervised machine learning clustering has shown promise in a number of music analysis applications, most notably in the field of song lyrics. The experiments discussed here demonstrate its uses in mood prediction, theme grouping, emotion detection, and genre categorization. These works show how machine learning methods may be used to comprehend the intricate and varied realm of music. More studies in this area are probably going to provide more profound understandings of the complicated connections between language, music, and human emotion as technology develops and more sophisticated datasets become accessible [26] [27].

3. Experimental work and its Evaluations

The experimental work and its evaluations are employed in this research paper is designed to rigorously analyze CokeStudio songs based on their lyrical content, utilizing a combination of natural language processing (NLP) techniques, topic modelling, and machine learning algorithms for clustering and classification. This section outlines the steps and procedures followed are discuss below:

3.1. Pre-processing and Data Collection

This section outlines the procedure for gathering and getting ready the data for analysis. The primary goals are to compile a complete dataset of CokeStudio songs and their lyrics and then do the necessary preprocessing on the data to make sure it is ready for further analysis.

Obtaining an extensive collection of CokeStudio song lyrics from several sources, including official websites and music streaming platforms, is the first stage in the process. The dataset, spanning from season 7 to season 14 and including songs from different cultures explored in season 2018, comprises approximately 282 songs. Only song transcripts or translations are included due to the songs being sung in multiple languages. Data preprocessing focuses mainly on the text/lyric column to standardize the text and remove noise and inconsistencies.

Key steps in pre-processing include:

- Text Tokenization: Splitting the text into individual words or tokens for further analysis.
- Stop-word Removal: To cut down on noise, remove frequently used terms that don't have much sense.
- Text Lowercasing: To guarantee consistency in analysis, every text should be converted to lowercase.
- Special Character Handling: Keeping some characters, such as hyphens and apostrophes, but removing punctuation and special characters that don't add anything to the analysis.
- Lemmatization/Stemming: To improve analysis, words might be truncated to their stem (stemming) or reduced to their base form (lemmatization).
- Handling Non-English and Regional Words: Considering only English translations of songs for uniform analysis, as Coke Studio songs often include non-English and regional terms.
- The collected dataset as shown below in the Figure 2.

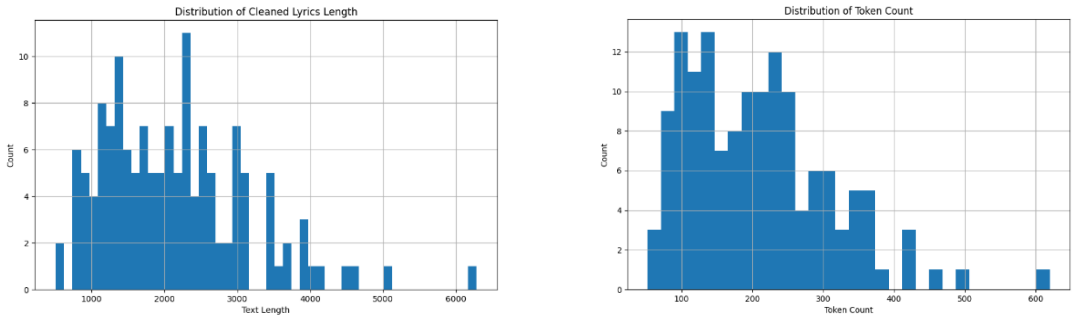
	Title	Singer	Text
0	O beautiful land	Various	O beautiful land, may God\nKeep you prosperous...
1	A Good Omen	Mai Dhai and Karam Abbas	I have brought a khair tree to plant in my hom...
2	For No Reason	Nabeel Shaukat Ali	O pain of parting, don't torment me for no rea...
3	O dear friends	Mekaal Hasan Band	Come together and congratulate me, my dear fri...
4	King of the Holy Sanctuary	Atif Aslam	Let a life of peace and contentment be my fate...
...	...	...	...
177	O Lover Who Has Left For Distant Lands	Zara Madani	O lover who has left for distant lands\nO love...
178	The Bird of Love	Sehar Gul Khan	The beloved qalandar arrives\n And knocks at t...
179	The Entire World	Bohemia	Listen to this tale of mine, my sweet\nthis ta...
180	(Raga Megh	Aziz Sohail	My beloved is away from me\nO friend, the nigh...
181	Inadvertently	Sanam Marvi	Ranjha\n Ranjha\n Ranjha\n Ranjha appeared as ...

182 rows × 3 columns

Figure 2. Collected dataset

4. Data Exploration

Following the completion of the data preparation, we conducted rudimentary exploratory analysis to obtain a deeper understanding of the dataset. This involves producing data such as the average amount of lyrics (around 209), the highest and minimum word counts per song (620 and 52, respectively), and the total number of songs (281). In Figure 3(a) and (b), the song distribution and any possible problems with the quality of the data are discussed, and some of the states are displayed.



(a) (b)

Figure 3. (a) and (b) show Distribution of Cleaned Lyrics length and Distribution of Token counts

5. Text Representation and Feature Extraction

This section focuses on converting pre-processed song lyrics into numerical representations for analysis and clustering purposes. While various techniques exist, including traditional bag-of-words models and advanced methods like word embedding's and topic modelling, the chosen method for vectorization is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF measures a word's importance in a song's lyrics relative to its occurrence across the entire dataset, highlighting distinctive words while balancing the impact of common ones. This approach captures semantic and contextual information, facilitating the identification of underlying themes and patterns in the lyrics. By utilizing TF-IDF, the lyrics are transformed into a numerical matrix, allowing for meaningful analysis and clustering while preserving the significance of words within each song and their importance across the entire corpus.

5.1. Topic Modelling

Finding underlying subject structures and trends in a vast collection of texts, like the lyrics of CokeStudio songs, may be accomplished with the use of a strong approach called topic modelling. Automatically identifying topics—groups of words that regularly co-occur in the text and indicate certain themes, thoughts, or ideas—is the main objective of topic modelling. These subjects shed light on the meanings, feelings, and stories that are conveyed via the lyrics of the songs.

We examine the vital steps involved in text representation and theme modelling as they relate to the lyrics analysis of CokeStudio songs. The three models, namely Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA), will be examined and their efficacy in identifying significant theme subjects from the poetic text will be assessed. We may give the subjects meaningful titles or themes by looking at the most frequent terms and their probabilities within each topic. For instance, a topic may be tagged as "Romantic Love" if it includes terms like "love," "heart," and "romance," etc. We employ three different topic extraction methods in our CokeStudio song lyrics analysis: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF).

5.1.1. Non-Negative Matrix Factorization (NMF)

NMF, as described by Lee and Seung [11], decomposes the TF-IDF matrix into interpretable topics and word distributions while enforcing non-negativity constraints. This method ensures that all matrix values remain non-negative, enhancing topic interpretability. In our analysis NMF topic models generated the top words for the CokeStudio songs as shown in figure 4 in word count for each topic.

5.1.2. Latent Semantic Analysis (LSA)

Deerwester et al. [12] created LSA, which employs singular value decomposition (SVD) to capture latent semantic associations between words and texts. A dimensionality reduction method called Latent Semantic Analysis (LSA) reveals latent theme elements in the lyrics of Coke Studio songs. Figure 5 displays the topics produced by LSA for each subject in the word count visualization.

Figure 4. Words count for each topic generated by NMF topic model

Figure 5. Word count visualization for each topic generated by LSA topic model

5.1.3. Latent Dirichlet Allocation (LDA)

The probabilistic generative model called LDA was presented by Blei et al. [13]. It works on the assumption that each text is a combination of themes, and that each topic is a distribution over words. LDA can capture the complex character of lyrical ideas since it generates topics in a flexible manner by seeing words as probabilistic entities.

5.1.4. Model Evaluation and Selection

The process of model evaluation and selection in our analysis of CokeStudio song lyrics involves assessing the coherence scores of topic modeling techniques like NMF, LSA, and LDA. Coherence scores indicate the quality and interpretability of topics generated by these models, helping us choose the most effective one.

Based on coherence scores Foltz,P W., et al. [14], NMF appears to perform the best, particularly with six topics, achieving a coherence score of 0.4090 as shown in Table 1. However, further exploration and visualization of topics are shown in figure 6.

Upon deeper analysis, NMF reveals six coherent topics, each represented by a collection of top words. These topics are labeled based on their overarching themes, such as Existence, Yearning, Devotion, Longing, Friendship, and Celebration as shown in figure 7. Assigning labels to topics facilitates song classification and clustering, enabling a structured analysis of CokeStudio song lyrics.

This approach provides insights into the diverse themes and patterns present in the songs, enhancing our understanding of the lyrical content.

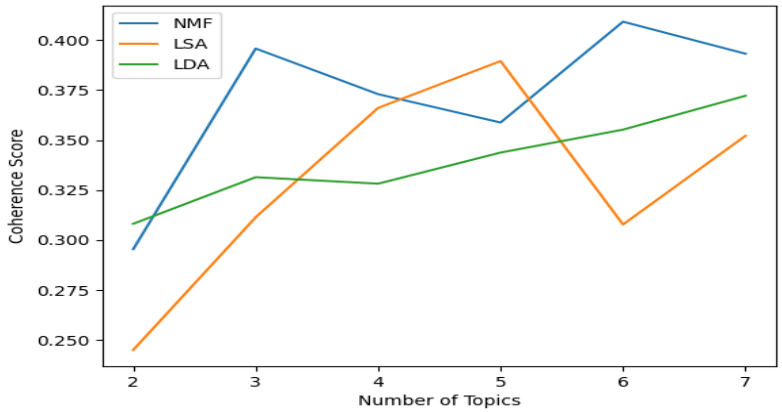


Figure 6. NMF, LSA and LDA Topic Models and Coherence score

Table 1. NMF, LSA, and LDA models Coherence Score

Topic No.	NMF Coherence Score	LSA Coherence Score	LDA Coherence Score
2	0.2954	0.2450	0.3081
3	0.3958	0.3114	0.3314
4	0.3730	0.3661	0.3282
5	0.3588	0.3895	0.3437
6	0.4092	0.3078	0.3552
7	0.3932	0.3521	0.3721
8	0.3407	0.3437	0.3309
Range	**(0.2954 to 0.4090)**	**(0.2450 to 0.3895)**	**(0.3081 to 0.3721)**
Highest Coherence Score	**0.4090**	**0.3895**	**0.3721**

Topic	Collection of Top word created by NMF	Label
Topic 1	life, world, beautiful, made, listen, God, nothing, leave, everything, even, turn, away, day, lord, last	**Existence**
Topic 2	beloved, home, sweetheart, eye, without, ranjha, heart, Allah, filled, lord, yearning, seen, colour, pas, wait	**Yearning**
Topic 3	master, Ali, God, breath, every, name, blessing, light, lord, present, lost, journey, path, lover, prayer	**Devotion**
Topic 4	heart, become, away, eye, thought, mine, lover, doe, memory, break, close, waiting, hope, tear, longing	**Longing**
Topic 5	friend, eye, mine, dear, wherever, together, became, speak, keep, name, tear, neither, without, end, peace	**Friendship**
Topic 6	dance, darling, mother, brother, sing, sorrow, joy, body, hair, head, happiness, play, away, dear, leave	**Celebration**

Figure7. Collection of words created by NMF model

5.2. Dimensionality Reduction

Computationally demanding high-dimensional representations may result in the "curse of dimensionality." Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) are two methods that can help with analysis and visualisation by reducing the dimensions while maintaining the most crucial information.t-SNE (t-Distributed Stochastic Neighbour Embedding): The dimensionality reduction method known as t-SNE was employed in our study to visualise high-

dimensional data derived from the lyrics of CokeStudio songs, with topics being determined using Non-Negative Matrix Factorizations (NMF). Because of its emphasis on maintaining the connections between data pieces, it may be used to identify patterns and clusters within the music. The labelled t-SNE method visualisation is shown in Figure 7.

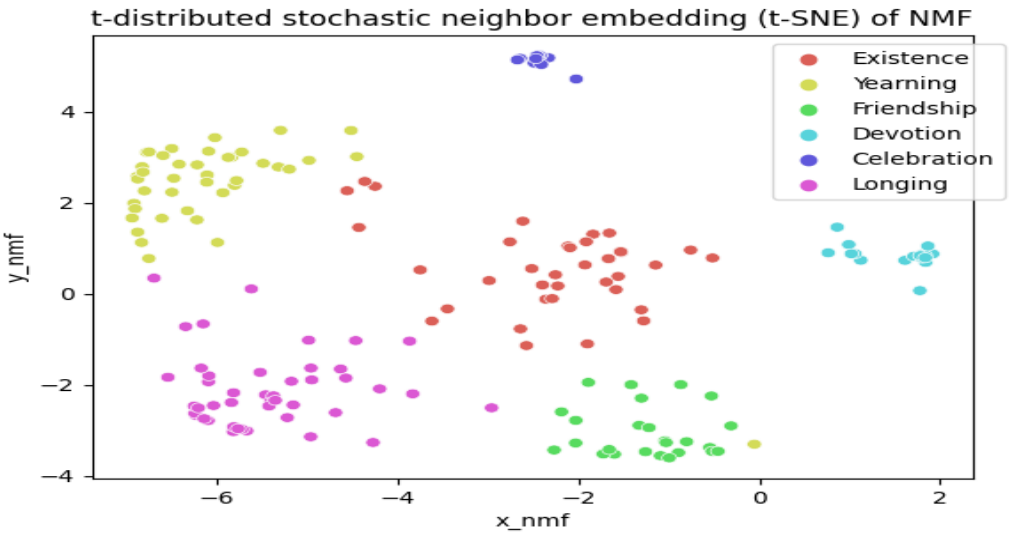


Figure 8. t-distributed stochastic neighbour embedding (t-SNE) of NMF values

6. Principal Component Analysis (PCA)

Using PCA, a dimensionality reduction technique, we were able to minimise the complexity of the CokeStudio song data while maintaining a high degree of variation in our analysis. It facilitates the discovery of underlying themes and content structures by enabling us to spot dominant patterns and elements in the lyrics. The PCA method's labelled visualisation is seen in Figure 8.

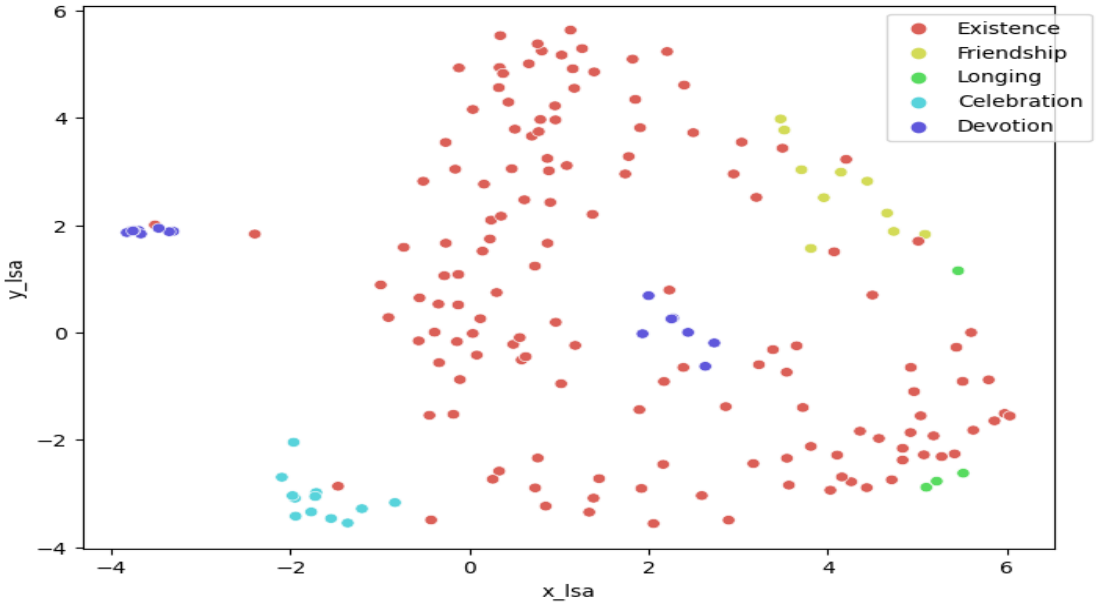


Figure 9. Principal Component Analysis (PCA) of LSA values

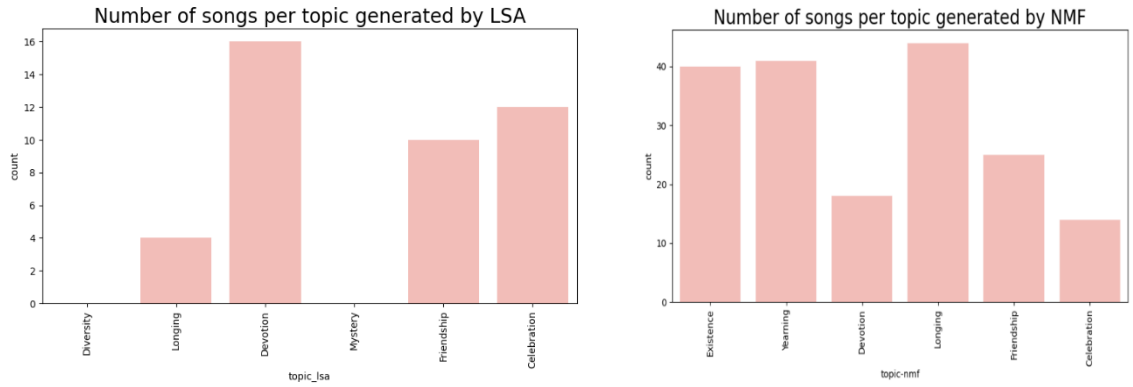
6.1. Songs per topic

The distribution of CokeStudio songs across several subjects produced by topic modelling Latent Dirichlet Allocation (LSA) is displayed in Figure 9(a) together with the Songs per topic visualisation (e.g., Diversity, Longing, Celebration, etc.). Additionally, the songs are visualised by topic using the Non-Negative Matrix Factorizations (NMF) topic model, as seen in Figure 9(b). It gives us a glimpse of the songs' thematic groupings and the variety of subjects found in the CokeStudio repertory.

As we can observe from Figure 9(a) and Figure 9(b) of our CokeStudio song lyrics dataset, the topics Devotion and Yearning have more songs created by the LSA topics model while Longing and Yearning have more songs generated by the NMF topic model, respectively.

7. CokeStudio Songs and Clustering

The goal of the machine learning approach known as clustering is to classify or cluster data items according to their intrinsic similarities. Our research aims to identify latent themes, patterns, or similarities between CokeStudio songs by clustering those together [35]. This aids in our comprehension of the many of subjects and genres found in the CokeStudio repertory. In order to cluster the CokeStudio Songs, we used clustering algorithms including K-Means, Agglomerative Clustering, and DBSCAN in our research. The descriptions of each clustering technique are provided below.



(a) (b)

Figure10. (a) and (b) show Number of songs per topic generated by NMF and LSA topic models

7.1. Optimal Number of Clusters Selection

Two different types of methods are used to choose the optimal number of clusters for the dataset we wish to locate clusters for. We'll investigate each one of them.

Silhouette Score: When choosing the ideal number of clusters for our study of CokeStudio songs based on lyrics/text, the silhouette score is a useful statistic. By comparing the similarity of data points inside a cluster to those of other clusters, it aids in the evaluation of the clustering's quality. An explanation of the silhouette score and its implications for our analysis is provided below: The cohesion and spacing between clusters are measured by the silhouette score. It falls between -1 and 1: There is a well-defined and adequate number of clusters when the score is near to 1, which indicates that the data points within a cluster are substantially similar to each other and distinct to those in other clusters. When data points on the border between clusters are unclearly allocated to one or the other, it is suggested that the clusters overlap (scores around 0). The data points are either improperly grouped or the number of clusters selected is unsuitable, as indicated by a score near to -1.

Results: Table 3 and Figure 10 demonstrate that, for n_clusters = 5, we have gotten the maximum silhouette score of almost 0.4164. This indicates that, among the cluster numbers that were examined, five clusters are the most suitable for categorising the CokeStudio songs according to their lyrics.

We must choose five clusters based on the silhouette score of 0.4164, which offers the optimal compromise between similarity within clusters and dissimilarity across clusters.

Table 2. Number of cluster and Silhouette Score

Number of Cluster	Silhouette Score
N_clusters = 2	0.3625
N_clusters = 3	0.3986
N_clusters = 4	0.3994
N_clusters = 5	0.4164
N_clusters = 6	0.3866
N_clusters = 7	0.3893

N_clusters = 8	0.3924
N_clusters = 9	0.3605

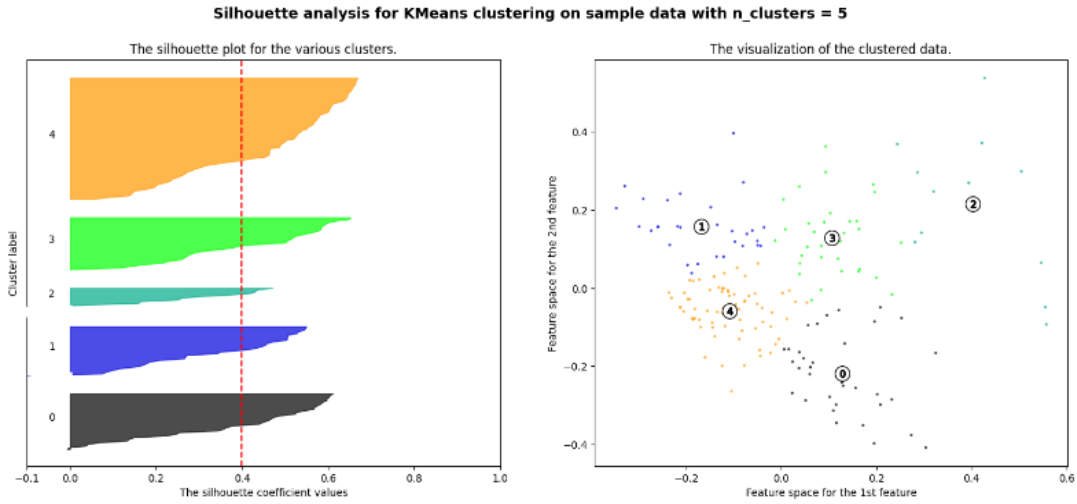


Figure 11. Clusters and Score of Silhouette

7.2. Elbow Method

To determine the ideal number of clusters given a dataset, the Elbow Method is a heuristic technique used in clustering research. Its foundation is the idea that as the number of clusters increases, the variation within each cluster diminishes, resulting in tighter, more cohesive clusters. There is a limit to how many clusters may be added before the variance is no longer appreciably reduced, leading to diminishing returns. Our dataset of CokeStudio songs was subjected to the Elbow Method, which determined the variance explained by various numbers of clusters.

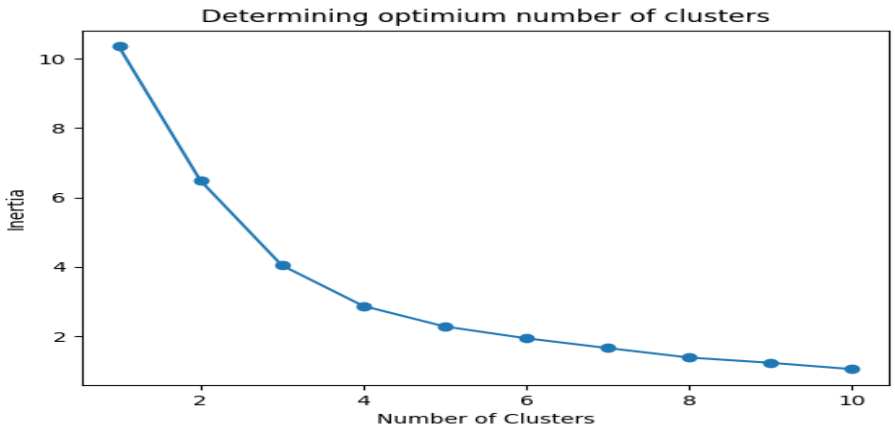


Figure 12. Selection of clusters

However, the elbow of the curve created on 5 indicates that 5 is the ideal number for clustering the lyrics dataset for our CokeStudio song, as Figure 11 illustrates.

7.3. K-means Clustering

K-means clustering is based on centroid clustering. The dataset is divided into K clusters by this method, and each cluster is represented by its centroid, or centre point. Based on the lyrical content, we have grouped CokeStudio tracks into K clusters using K-means. K-means finds the cluster centroids that most accurately capture each group's key themes by taking into account the TF-IDF representations of song lyrics. Every K-means cluster consists of a collection of songs with similar lyrical topics. This might aid in comprehending the many CokeStudio song classifications. Figure 12 depicts the depiction of the clusters to acquire insights into how CokeStudio songs are spread across different lyrical themes or genres.

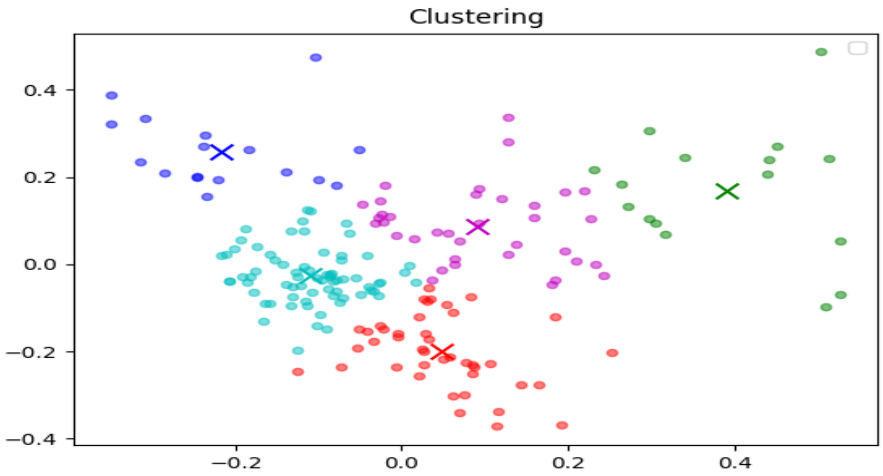


Figure 13. Clusters formation using K-mean Algorithm

7.4. Agglomerative Clustering

One type of hierarchical clustering method is agglomerative clustering. A hierarchy of clusters is created by iteratively combining individual data points at the beginning to form clusters. Clusters' hierarchical structure enables you to investigate several lyrical similarity levels, from overarching themes to more specific sub-themes. We created a hierarchical representation of CokeStudio tracks using a technique called agglomerative clustering. It facilitates our comprehension of the connections between songs at various granularities. To obtain an understanding of the distribution of CokeStudio songs across various lyrical topics or genres, we have visualised the clusters in Figure 13.

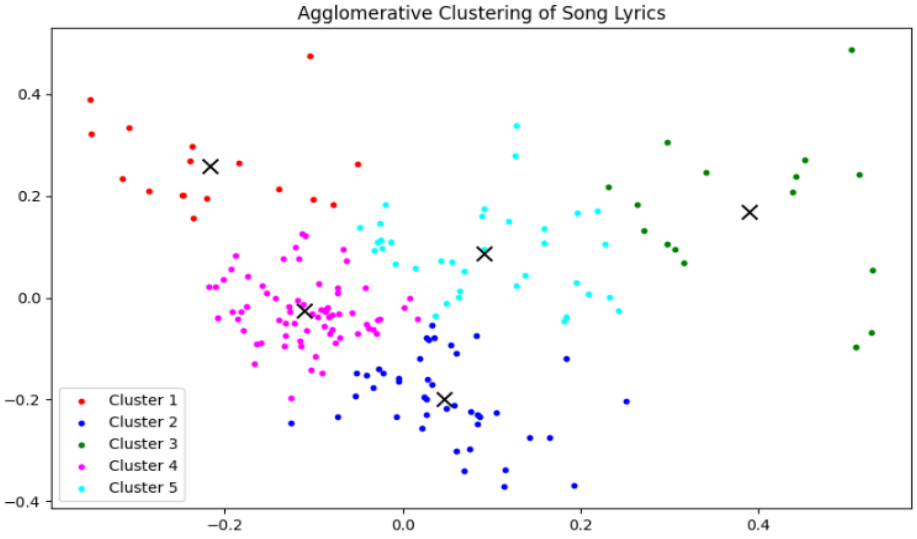


Figure 14. Clusters formation using Agglomerative Algorithm

7.5. DBSCAN

Since DBSCAN is a density-based clustering method, it can distinguish between areas with high and low data densities and clusters nearby data points in the feature space. Zhang, M. DBSCAN is appropriate for datasets where the number of clusters is unknown in advance since, unlike K-means, it does not need you to define the number of clusters beforehand. In order to find groupings of songs with related lyrical themes or substance, we used the DBSCAN clustering technique to cluster CokeStudio tracks based on their lyrics or text. The CokeStudio Songs Clustering and Visualisation may be achieved using DBSCAN in the following ways, as seen in Figure 14:

- No Preset Number of Clusters: DBSCAN's capacity to identify clusters without specifying the number of clusters ahead of time is useful because we do not know in advance how many unique lyrical themes or song categories are present in the CokeStudio dataset.
- Managing Noise: DBSCAN is able to recognise and classify songs as noise points that have odd or loud lyrics. This aids in removing tracks that don't belong in any cluster.

- Density-Based Clusters: DBSCAN works effectively for locating groups of songs that differ in size and form. It can identify both sparse and densely packed clusters, which may indicate various lyrical patterns in songs produced using CokeStudio.
- Interpretable Results: We are able to examine and comprehend the variety of songs in the dataset by interpreting the clusters produced by DBSCAN in terms of lyrical topics.

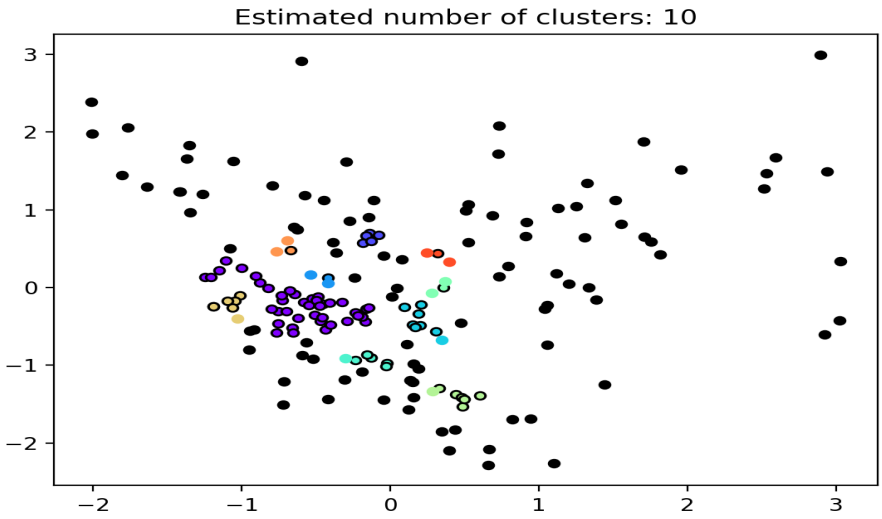


Figure 15. Clusters formation using DBSCAN

Following the use of DBSCAN, the algorithm begins by selecting a random data point and growing the cluster by appending all accessible core points to it (as seen in Figure 4.23, we have added a circle to each point). Until no more core points can be added to the cluster, this procedure keeps on. The programme then chooses a different, unseen data point and goes through the same steps again, adding new clusters or labelling noise points along the way. DBSCAN makes sure that our clusters can adjust to the density of lyrical content in the data and are not just based on a fixed number (as in K-means). Ten clusters are estimated using DBSCAN for our CokeStudio music dataset.

7.6. K-Means-determined Most Common Word for Each Cluster

In order to identify unique themes within the songs, we have finished applying K-Means clustering on the CokeStudio song lyrics dataset. Following the application of K-Means clustering with five clusters, we examined the resulting clusters and determined which terms were most often found inside each cluster. An overview of the most often used terms in each cluster is provided below:

7.6.1. Universal Existence, Cluster 1

As seen in Figure 15(a), this cluster incorporates the universal concepts of existence and life. It investigates the intricacies of our reality and the beauty inherent in the ordinary. The terms "life," "world," "even," and "everything" convey an awareness of the complexity of the environment and a range of experiences.

7.6.2. Cluster 2: Love and desire

Cluster 2 explores the depths of love and desire (Figure 15(b)). There are several emotional terms in this cluster, including "beloved," "heart," and "yearning." These words probably convey intense emotions and a yearning for someone or something unique.

7.6.3. Beauty and Dance Cluster 3

Beauty and festivity are radiated from this cluster. Words like "beautiful," "dance," "colour," and "darling" describe it. It stands for the good times in life and the appreciation of beauty in all its manifestations. Moreover, figure 16 (a) illustrates.

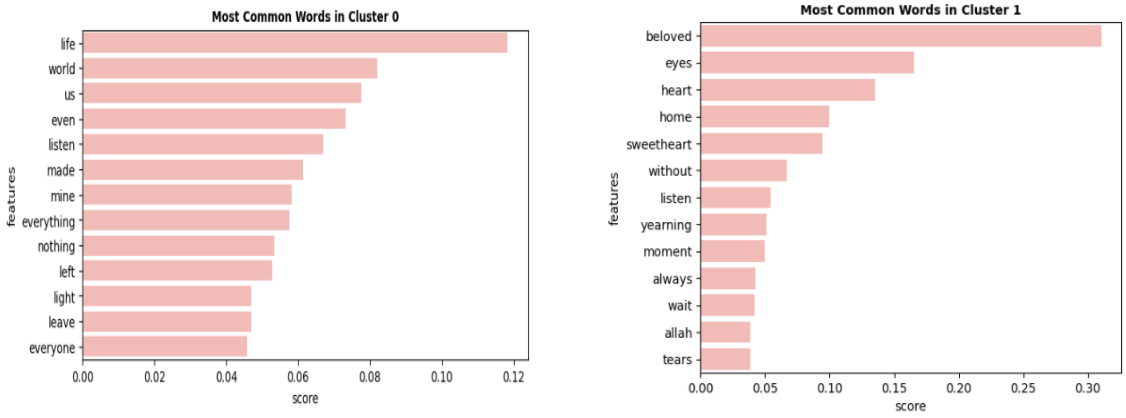
7.6.4. Cluster 4: Hope and Longing:

As seen in figure 16(b), Cluster 4 is dominated by hope and longing. It depicts the feelings of becoming, waiting, and longing. Words that convey anticipation and a want for a better future include "heart," "waiting," and "hope."

7.6.5. Cluster 5: Spirituality and Devotion

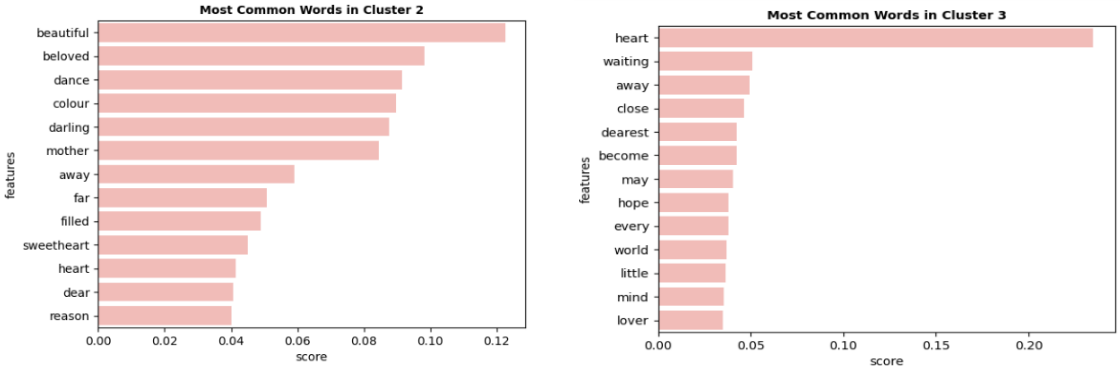
As seen in Figure 17, Cluster 5 has a spiritual and devotional element. Words like "god," "master," "ali," and "lord" are among them. The lyrics in this cluster may allude to a strong spiritual feeling and a close

relationship with the divine. The fundamental feelings and ideas that the CokeStudio songs within each cluster are trying to portray are made clearer to us by these theme interpretations.



(a) (b)

Figure 16. (a) and (b) Show Most Common Words of Cluster 1st and 2nd using k-mean



(a) (b)

Figure17. (a) and (b) Shows Most Common Words in Cluster 3 and 4 of k-mean

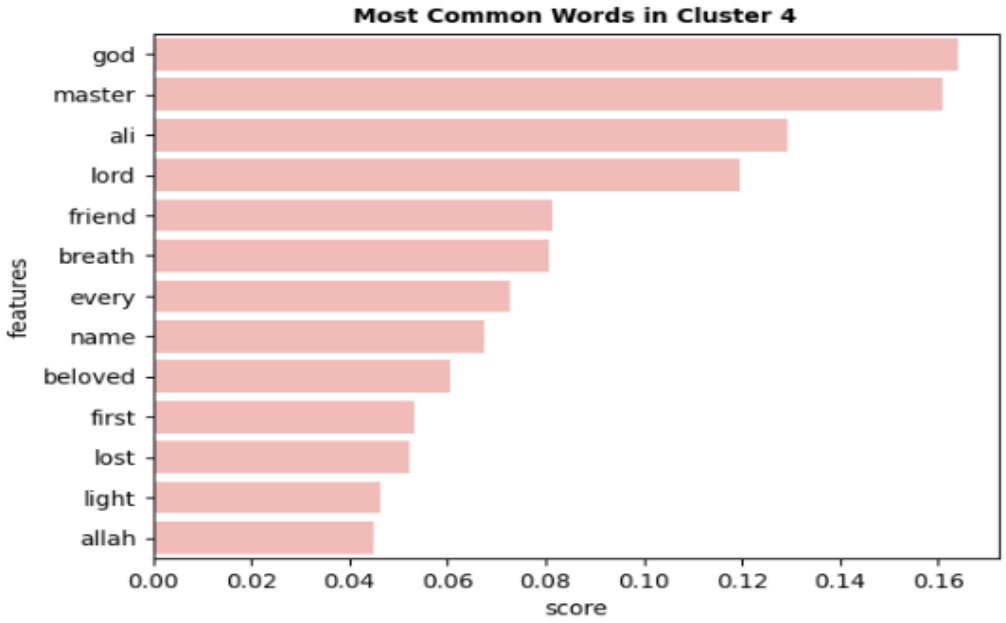


Figure 18. Most Common Words of Cluster 5 using k-mean

They make it simpler to examine the creative and emotional depth of the songs in our CokeStudio Songs collection by offering an organised method for classifying and analysing the lyrics based on the recurrent themes.

8. Classification of CokeStudio Songs

Assigning data points to predetermined groups or classes is the goal of the supervised machine learning job of classification. In this case, we investigated a number of classification models employing topic labels produced by NMF (Non-Negative Matrix Factoriza-tion) and LSA (Latent Semantic Analysis) models to group CokeStudio songs according to their lyrical content. A number of algorithms were examined, and their efficacy in our clusters was assessed based on training and testing scores. These algorithms included Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Multinomial Naive Bayes.

Logistic Regression, for instance, showed strong performance with high training and testing scores for both NMF and LSA features, making it a reliable choice for classifying song topics. Decision Trees, while achieving perfect training scores, had lower testing accuracy, especially with NMF-labeled data. Random Forest, an ensemble method, also performed well but had a noticeable drop in testing accuracy. Gradient Boosting, known for its high predictive accuracy, demonstrated significant effectiveness, particularly with NMF values, achieving the highest testing scores among the models. Lastly, Multinomial Naive Bayes, a probabilistic classifier, effectively modeled the probability distribution of words, aiding in thematic classification of songs. Each model's performance highlights the varying strengths and suitability for thematic classification tasks in textual data analysis.

Let's review the models we evaluated and then decide which seems to be the best based on the ratings that were given, which are displayed in Table 4.

Table 3. NMF and LSA Testing and Training Score

Classification Algorithm	For NMF Topic Model		For LSA Topic Model	
	Training Score	**Testing Score**	**Training Score**	**Testing Score**
Logistic Regression	0.9917	0.8948	0.9969	0.8421
Decision Tree	1.0	0.5263	1.0	0.6842
Random Forest	1.0	0.8421	0.838	0.473
Gradient Boosting	0.3958	0.2105	0.80	0.6316
Multinomial Classification	0.9792	0.7368	0.9748	0.6842

8.1. Classification reports

Table 5 of the classification report offers a thorough assessment of how well the Logistic Regression model performed in classifying CokeStudio tracks according to their lyrical content.

Table 4. Logistic Regression Classification Report

Classes	Precision	Recall	F1-Score
0	1.00	1.00	1.00
1	0.80	1.00	0.89
2	1.00	1.00	1.00
3	1.00	1.00	1.00
4	0.67	0.67	0.67
5	1.00	0.67	0.80
Accuracy			**0.89**
Macro avg	**0.92**	**0.89**	**0.89**
Weighted avg	**0.91**	**0.89**	**0.89**

8.2. Confusion Matrix

As seen in Figure 4.18, a confusion matrix aids in our evaluation of how effectively our model is able to discern between various CokeStudio song categories based on their lyrics. It's a useful tool for assessing how well our categorization strategy is working and pinpointing areas that need work.

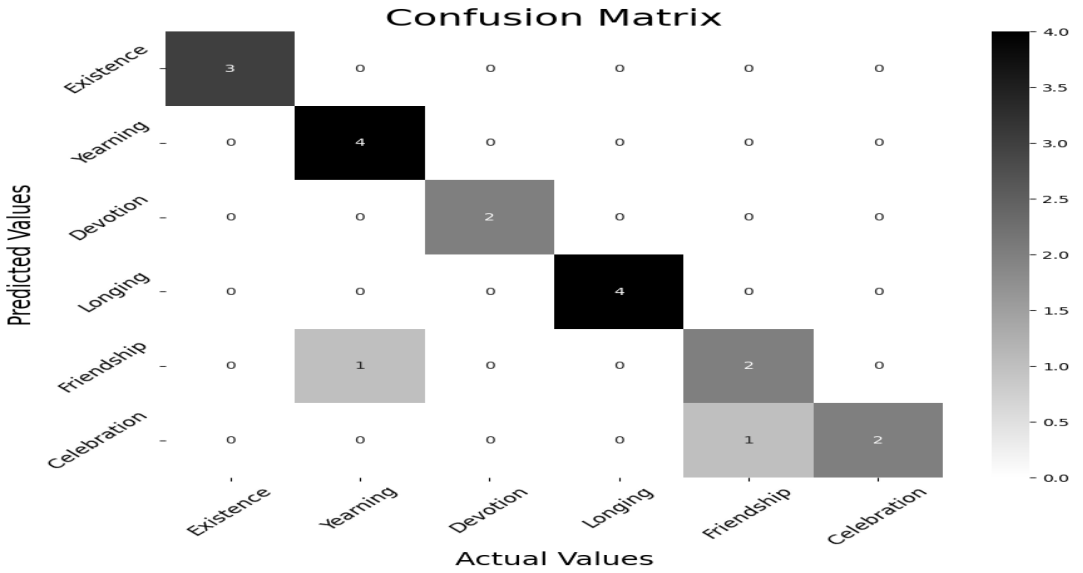


Figure 19. LR Confusion Matrix

9. Conclusions

We used a variety of data analytic tools to perform a thorough examination of CokeStudio songs, concentrating on their lyrical content. This involved preprocessing the data, using feature engineering to convert unstructured text data from the lyrics, and using topic modelling techniques to identify six main themes in the songs. In the end, we used English transcripts for consistency while also looking at the length and intricacy of the lyrics as well as cultural and linguistic considerations. We also performed clustering and classification analysis. We grouped songs according to lyrical similarities using clustering methods such as K-Means, Agglomerative Clustering, and DBSCAN, which revealed different themes. Additionally, we employed supervised machine learning methods, including Multinomial Naive Bayes, Random Forest, Decision Trees, Gradient Boosting, and Logistic Regression with SMOTE applied to NMF data. The results showed that this method produced the best accuracy, at 89.5%. With possible applications in the future for content recommendation systems and more in-depth cultural study across CokeStudio's extensive music catalogue, this research is practically significant for automated classification and subject exploration. The approach may be modified to analyse music from different languages and cultural backgrounds.

References

1. Abid, S.; Bilal, M. Z.; and Begum, A. "Music and Trans culturalism: analyzing the role of coke studio music in Pakistan", 2022, Pakistan Journal of Social Research, vol. 4, no. 03, pp. 933-943.
2. Yang, D.; and Lee, W. S. "Music emotion identification from lyrics", 2009, In 2009 11th IEEE International Symposium on Multimedia (pp. 624-629). IEEE.
3. Ali, M. A.; and Siddiqui, Z. A. "Automatic music genres classification using machine learning", 2021, International Journal of Advanced Computer Science and Applications, vol. 8, no. 8, pp. 201-216.
4. Siddique, M. A. S.; Sarker, M. I.; Ghosh, R.; and Gosh, K. "Toxicity Classification on Music Lyrics Using Machine Learning Algorithms", 2021, In 2021 24th International Conference on Computer and Information Technology (ICCIT), (pp. 1-5). IEEE.
5. Nandwani, P.; and Verma, R. "A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining", 2021, vol. 11, no. 1, pp. 81-95.
6. Rarasati, D. B. "A grouping of song-lyric themes using k-means clustering", 2020, JISA (Jurnal Informatika dan Sains), vol. 3, no. 2, pp. 38-41.
7. Lee, D. D.; and Seung, H. S. "Learning the parts of objects by non-negative matrix factorization" 1999, Nature, vol. 401, no. 6755, pp. 788-791.
8. Blei, D. M.; Ng, A. Y.; and Jordan, M. I. "Latent dirichlet allocation", 2003, Journal of machine Learning research, pp. 993-1022.
9. Foltz, P.; Kintsch, W.; and Landauer, T. K. "The measurement of textual coherence with latent semantic analysis", 1998, Discourse processes, vol. 25, no. 2, pp. 285-307.
10. Zhang; M. "Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members", 2019, IOP Conference Series: Earth and Environmental Science, 252, 042033. doi:10.1088/1755-1315/252/4/042033
11. Khalil, M. I. K.; Afsheen, A.; Taj, A.; Nawaz, A.; Jan, N.; and S. Ahmad, "Enhancing Security Testing Through Evolutionary Techniques: A Novel Model" Journal of Computing & Biomedical Informatics, 2023, vol. 6, no. 1, pp. 375 – 393.
12. Khalil, M. I. K. and Taj, T. "Factors Affecting the Efficacy of Software Development: A Study of Software Houses in Peshawar, Pakistan," International Review of Basic and Applied Sciences, 2021, vol. 9, no. 3, pp. 385-393.
13. Khalil, M. I. K. Ullah, A.; Taj, A.; Khan, I.A.; Ullah, F.; Taj, F.; and Shah, S. "Analysis of Critical Risk Factors Affecting Software Quality: A Study of KPK, Pakistan Software Industry," International Review of Basic and Applied Sciences, 2022, vol. 10, no. 2, pp. 338-348.
14. Saqib, A.; Ullah, M.; Hyder, S.; Khatoon, S.; and Khalil, M. I. K. "Creative Decision Making in Leaders: A Case of Beer Game Simulation," Abasyn Journal of Social Sciences, 2020, vol. 12, no. 2, pp. 379-387.
15. Khan, I.A; Ullah, F.; Abrar, M.; Shah, S.; Taj, F. and Khalil, M.I.K. "Ransomware Early Detection Model using API-Calls at Runtime by Random Decision Forests," International Review of Basic and Applied Sciences, 2022, vol. 10, no. 2, pp. 349-359.
16. Jan, S.; Maqsood, I.; Ahmad, I.; Ashraf, M.; Khan, F. Q.; and Khalil, M. I. K. "A Systematic Feasibility Analysis of User Interfaces for Illiterate Users," Proceedings of the Pakistan Academy of Sciences, 2020, vol. 56, no. 4, pp. 2518-4253.
17. Khalil, M. I. K.; Shah, S. A. A.; Khan, I. A.; Hijji, M.; Shiraz, M.; and Shaheen, Q. "Energy cost minimization using string matching algorithm in geo-distributed data centers," Computers, Materials, and Continua, 2023, vol. 75, no. 3, pp. 6305-6322.
18. Ahmad, I.; Khalil, M. I. K.; and Shah, S. A. A. "Optimization-based workload distribution in geographically distributed data centers: A survey," International Journal of Communication Systems, 2020, vol. 33, no. 12, p. e4453.
19. Khalil, M. I. K.; Ahmad, I.; Shah, S. A. A.; Jan, S.; and Khan, F. Q. "Energy cost minimization for sustainable cloud computing using option pricing," Sustainable Cities and Society, 2020, vol. 63, p. 102440.
20. Khalil, M. I. K. "Improve quality of service and secure communication in mobile adhoc networks (MANETS) through group key management," International Review of Basic and Applied Sciences, 2013, vol. 1, no. 3, pp. 107-115.
21. Muhammad, D.; Ahmad, I.; Khalil, M. I. K.; Khalil, W.; and Ahmad, O. A. "A generalized deep learning approach to seismic activity prediction," Applied Sciences, MDPI, 2023, vol. 13, p. 1698.
22. Khalil, M. I. K. "Job satisfaction and work morale among Ph.D's: A study of public and private sector universities of Peshawar, Pakistan," International Review of Management and Business Research, 2013, vol. 02, no. 2, p. 362.
23. Ahmad, I.; Ahmad, M. O.; Alqarni, M. A.; Almazroi, A. A.; and Khalil, M. I. K. "Using algorithmic trading to analyze short-term profitability of Bitcoin," PeerJ Computer Science, 2021, vol. 7, p. e337.
24. Khalil, M. I. K.; Shah, S. A. A.; Taj, A.; Shiraz, M.; Alamri, B.; Murawat, S.; and Hafeez, G. "Renewable aware geographical load balancing using option pricing for energy cost minimization in data centers," Processes, MDPI, 2022, vol. 10, no. 10, p. 1983.
25. Khalil, M. I. K.; Ahmad, A.; Almazroi, A.A. "Energy Efficient Workload Distribution in Geographically Distributed Data Centers," IEEE Access, 2019, vol. 7, no. 1, pp. 82672-82680.
26. Naz, S., Amin, H., & Sayed, A. (2024). Maternal Mortality in Pakistan: The Potential Role of Community Midwives. Journal of Development and Social Sciences, 5(2), 45-52.
27. Naz, S., Aslam, M., & Sayed, A. (2023). Prevalence of Anemia and its Determinants among the Rural Women of Khyber Pakhtunkhwa-Pakistan. Annals of Human and Social Sciences, 4(4), 42-50.

28. Khalil, M. I. K..; Mubeen, A.; Taj, A.; Jan, N.; Ahmad, S. "Renewable and Temperature Aware Load Balancing for Energy Cost Minimization in Data Centers: A Study of BRT, Peshawar," Journal of Computing & Biomedical Informatics, 2023, vol. 06, no. 3, pp. 1-8.
29. Hussain, S.K., Ramay, S.A., Shaheer, H., Abbas T., Mushtaq M.A., Paracha, S., & Saeed, N. (2024). Automated Classification of Ophthalmic Disorders Using Color Fundus Images, Volume: 12, No: 4, pp. 1344-1348 DOI:10.53555/ks.v12i4.3153
30. Anum, H.; Khalil, M. I. K..; Nawaz, A.; Jan, N.; and Ahmad, S. "Enhancing rumor detection on social media using machine learning and empath features," Journal of Computing & Biomedical Informatics, 2023, vol. 06, no. 2, pp. 6-13.
31. Khalil, M. I. K.; Khan, I.A; Nawaz, A.; Latif, S.; Ahmad, S. and Ahmad, S. "Unveiling the security Maze: A comprehensive review of challenges in Internet of things," Special Issue on Intelligent Computing of Applied Sciences and Emerging Trends (ICASET), Journal of Computing & Biomedical Informatics, 2024, pp. 10-19.
32. Ahmad, I.; Ahmad, M. O.; Alqarni, M. A.; Almazroi, A. A.; and Khalil, M. I. K. "Using algorithmic trading to analyze short-term profitability of Bitcoin," PeerJ Computer Science, 2021, vol. 7, p. e337.
33. Khalil, M. I. K.; Shah, S. A. A.; Taj, A.; Shiraz, M.; Alamri, B.; Murawat, S.; and Hafeez, G. "Renewable aware geographical load balancing using option pricing for energy cost minimization in data centers," Processes, MDPI, 2022, vol. 10, no. 10, p. 1983.
34. Abbas, M., Arslan, M., Bhatty, R. A., Yousaf, F., Khan, A. A., & Rafay, A. (2024). Enhanced Skin Disease Diagnosis through Convolutional Neural Networks and Data Augmentation Techniques. Journal of Computing & Biomedical Informatics, 7(01).
35. Zaidi, A., Karim, A. A., Mohiuddin, S., Khan, A., Syed, A., Jehangir, M., & Afzal, I. (2018). Dental Sensitivity Associated With Consumption Of Fizzy Drinks: A Cross Sectional Study. Pakistan Journal of Medicine and Dentistry, 7(4), 5-5.