Viral Melodies: Exploring the Factors Influencing Music Virality in TikTok Engagement

1st Evelyn Zefanya Rahardjo
Data Science Program
Computer Science Department
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
evelyn.rahardjo@binus.ac.id

2st Jasmine Mutia Alifa
Data Science Program
Computer Science Department
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
jasmine.alifa@binus.ac.id

3st Sharon Zefanya Setiawan
Data Science Program
Computer Science Department
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
sharon.setiawan001@gmail.com

4st Alexander Agung Santoso Gunawan
Computer Science Department
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
aagung@binus.edu

5st Karli Eka Setiawan
Computer Science Department
School of Computer Science
Bina Nusantara University,
Jakarta, Indonesia 11480
karli.setiawan@binus.ac.id

***Abstract*—In the digital era and the evolution of social media platforms like TikTok, understanding the factors influencing content virality has become increasingly crucial. Therefore, this research aims to delve into the musical variables contributing to the popularity of content on TikTok, including aspects such as artist, beat of music, and the number of shares. Focusing on content that successfully garnered over 5,000 likes, this study employed a data analysis approach involving the librosa library for music tempo extraction and logistic regression to identify relationships among these variables. The research findings indicate that the most significant variable influencing the virality of music content on TikTok is the number of shares. This highlights that the phenomenon of virality is not solely dependent on musical characteristics like artist or beat but is more influenced by social interactions through video sharing among users. The findings suggest that when someone shares a video with others who may have similar music preferences, the video is likely to receive likes, even if it does not appear on the user's For You Page (FYP). Thus, the virality of content on TikTok can be explained by the existence of social connections among friends that trigger the viral spread of the video. These results provide valuable insights into the mechanisms behind the success of conten/t on this platform.**

Keywords—tiktok, music selection, user engagement, content virality, data analysis

I. INTRODUCTION

TikTok is a famous social media site and has more than billions of users. TikTok has become a forum for content creators to express themselves and do business [1]. In creating content, music is an important element to encourage other users' interest in watching the content. Choosing the right music knowing the factors that make content go viral can trigger emotions and connect the audience and the content they watch [2].

Music on TikTok is an element for users to express themselves in a unique way. Music has the power to convey the emotions of its users and can create emotional bonds between people [3]. The success of content is also influenced by the choice of music. This research aims to explore musical variables that contribute to the popularity of content on TikTok, including aspects such as artist, musical rhythm, and number of shares that make this paper unique. This paper provides a new perspective on the impact of music culture on TikTok.

To gain deeper insights, Researchers used data from Kaggle to analyze the impact of music choices on the virality of TikTok content, providing insights for content creators and marketers. Descriptive analysis methods are used to understand the characteristics and patterns of data [5], with descriptive statistical techniques such as bar charts, pie charts, and many others providing a clearer understanding of the data.

This paper discusses the importance of understanding the virality of content on TikTok and explains what variables can contribute to the popularity and virality of content.With in-depth analysis, this paper provides useful insights for content creators to maximize TikTok's potential as a tool for self-expression.

II. RELATED WORKS

Several research studies were found related to this research topic. First, there is Research conducted by Olivia Sadler [6] shows that protest music on TikTok plays a role in movements for social change in America, amplifying voices on social issues and influencing emotions and engagement. Second, there is a paper by Wang Y [7] discussing the effect of musical rhythm on the listener's emotions. In this study, non-musical students listened to seven types of music, and their emotional reactions were measured using EEG. Results showed that almost all types of music influenced emotions, with electronic and light classical music having different emotional responses but similar levels of recognition.

Han Yang's research [8] explored the use of music in marketing on TikTok, finding that music increases attention and popularity. Terrence Cook [9] revealed that musical preferences influence emotional expression. These two studies confirm the important role of music in influencing emotions and engagement. Research conducted by the Institute of Musicology SASA, Belgrade, Serbia [10] explores the influence of music on TikTok, focusing on the process of creating, listening to, and promoting music. This study reveals how music increases engagement and sound integration in TikTok content. The paper researched by T. Shaikh [11] suggests that music genre classification using neural networks is very effective and has a high level of accuracy.

Another study conducted by Vizcaíno-Verdú A. [12]. He stated that users in the role of music curators preferred video formats on platforms like YouTube or TikTok over audio-only. Research conducted by Choi.K [13]. demonstrated the success of deep neural networks (DNNs) in music classification and tagging. Finally, Simarmata I.L [14] evaluated the use of Random Forest and XGB Classification models in music genre classification, with Random Forest achieving 72% accuracy and XGB Classification 73%, both effective despite different approaches.

III. RESEARCH METHODOLOGY

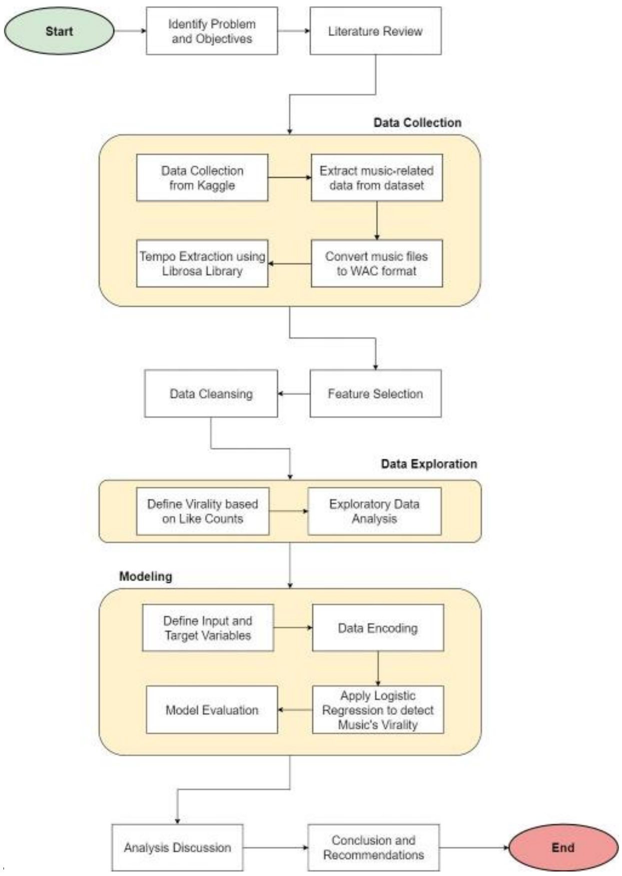

Fig 1. Research Flow Diagram

A. Data Collection

The dataset used TikTok Trending Videos comes from Kaggle [15] and the dataset includes audio datasets from Spotify and Apple Music in CSV format, videos from TikTok in MP4 format, and a collection of trending videos in JSON format.

The initial focus is on understanding the structure and types of data for quality analysis [16]. Kaggle datasets are used for automatic extraction of text, audio, user profiles, and other features, producing complete datasets for analysis. Since the raw data is unstructured so comprehensive pre-processing is essential, this includes music tempo detection and JSON data is our main data set. This raw data consists of 5693 records with 17 characteristics related to TikTok engagement. The following is a description of the json data:

TABLE 1. EXPLANATION OF EACH VARIABLE

Variable	Description
id	The unique identifier of the video
text	The text below of the video
createTime	Timestamp of the datetime when the video was created
authorMeta	An object with detailed information about the author
musicMeta	An object with detailed information about the music used with the video
covers	An object with all covers of the video
webVideoUrl	Link to the Tiktok Video
videoUrl	Exact link to the Tiktok video (not reachable directly)
videoUrlNoWaterMark	The URL of the video without a watermark
videoMeta	An object with dimensions and duration of the video
diggCount	Amount of likes
shareCount	How many times the video has been shared
playCount	How many times the video has been watched
commentCount	Amount of comments
downloaded	If the video is downloaded using the scraper
mentions	List with users mentioned in the video
hashtags	List with hashtags used in the video

The current challenge involves the extraction and integration of new datasets to enhance the analysis of complex relationships between music attributes and genre classifications in the context of TikTok engagement. Starting with 'trending.json' data, the step is to convert the audio file from URL to WAV format by taking the audio URL from the dataFrame df_tiktok_audd_music (which comes from the json file), download it using requests.get, and then convert the downloaded audio to WAV format using AudioSegment.from_file. After conversion, the audio is saved as a WAV file with a name that includes an index number. This process is performed for each row in the first 999 rows of the DataFrame, and if a download failure occurs or the URL is invalid, the code flags the failure. Finally, a new DataFrame containing the WAV file name and URL is created and saved as a CSV file. The purpose of converting MP3s to WAV files is to make analysis easier with the Librosa library for calculating beats.

Here the beat_track function is created. The beat_track function calculates the BPM of a WAV audio file using the librosa library (librosa.beat.beat_track), then classifies the song by tempo into "Slow", "Medium", "High", or "Unknown" if the BPM is not within a certain range. The following are the BPM ranges used for tempo classification:

- Fast: 20 - 69 bpm
- Medium: 70 - 109 bpm
- Slow: >= 110 bpm

Through this detailed procedure, the dataset undergoes a refinement that elevates its depth and precision, enabling a more nuanced exploration of the correlation between tempo attributes and musical genres in the TikTok engagement domain.

Similar to the previous steps, it is essential to remove unnecessary attributes before shaping the data for further processing. This pruning of attributes streamlines the dataset, ensuring that only the relevant information is retained for subsequent stages. Following this attribute refinement, a new dataset will be curated and saved in the CSV file format. The variables in the new dataset are the variables that have the strongest correlation with the virality of a video.The outcomes of the dataset extraction are as follows:

TABLE 2. EXPLANATION OF EACH EXTRACTED VARIABLE ⍰ VIRAL

Variable	Description
URL	Link to the audio files
webVideoUrl	Link to the Tiktok videos
Filename	Music in format WAV
artist	Artist of the audio tracks
title	Title of the audio tracks
likeCount	Amount of likes
shareCount	How many times the video has been shared
playCount	How many times the video has been watched
commentCount	Amount of comments
bpm	A measure of speed or tempo in music
beat	Beats or pulses in a musical composition

B. Data Preprocessing

In this research, Python is used for data analysis with tools such as pandas, numpy, re, and unidecode [18]. The initial stage involves data pre-processing [19] to address issues such as missing data, normalization, and outliers.

From a cursory glance, it appears that a significant portion of the data is textual and still needs cleansing. As examples, consider the artist and title column, which are characters, symbols, and foreign languages [20]. First things first: use the unidecode library to replace accent characters in text with equivalent characters without accents. Next, replace special characters and delete foreign characters, symbols, and extra spaces.

In social media data analysis, converting string data to numeric is critical to assessing the effectiveness of content or marketing initiatives, such as converting “likes,” “comments,” and “shares” to numeric values. This enables data-driven decisions and identifies trends that might be missed if data were stored as strings [21].

The data analysis highlights the need for refining the data types, particularly for metrics such as likes, shares, views, and comments, currently assigned as floating-point numbers. Recognizing the inherently quantitative nature of these metrics, the adjustment involves converting them to integers for a more accurate representation. This modification is crucial to maintain data integrity and ensure the appropriateness of numerical information.

Additionally, the analysis identifies issues with duplicate entries, particularly in the 'webVideoUrl' category, indicating redundancy in the dataset. To address this, a careful curation process is recommended to create a more refined subset, guaranteeing unique entries and resolving missing values. These steps contribute to enhancing the dataset's precision and completeness for subsequent analyses and applications.

C. Data Exploration (EDA)

The next stage is EDA. EDA aims to explore, understand the prepared data and obtain information from the data [22]. The goal of data exploration is to understand the distribution of data, produce visualizations, understand descriptive statistics, investigate relationships and trends, and identify important features.

The research employs a comprehensive analysis of TikTok content virality, utilizing statistics such as high averages and extreme values in likes and views. The classification into viral or non-viral categories is based on the 'likeCount,' resulting in the creation of a 'virality' column. This column is determined by the 'detect_virality' function, categorizing content as 'Viral' or 'Nonviral' based on the 'likeCount,' with 302 videos falling under the 'Viral' category.

This research combines Matplotlib and Seaborn to examine the correlation between music tempo and the popularity of TikTok content. Using bar graphs and pie charts, this research reveals the relationship between 'beats' and content virality. Next, the rhythm variations of the top 10 artists are analyzed, with Seaborn's visualization clarifying the interactions between variables through correlation matrices and heat maps. The goal is to understand how musical dynamics influence the success of content on TikTok. Finally add a 'shareGroup' column to the dataframe, create a 'detect_share' function to classify 'shareCount' as 'Low', 'Medium', or 'High'. The aim is to evaluate the influence of 'shareCount' on content virality. After EDA, the next step is to determine the independent and dependent variables.

D. *Determining Input and Output*

In logistic regression, determining the input (independent variable) and output (dependent variable) is an important step to understand the cause-and-effect relationship between two or more quantitative variables [23]. Inputs are factors that we believe influence the variables we observe or predict (output). The following is a table for determining input and output:

Variable	Description	Input / Output
artist	Artist of the audio tracks	Input → X
title	Title of the audio tracks	Input → X
shareCount	How many times the video has been shared	Input → X
virality	explanation of viral or non-viral video	Output (Target Variable) → y

E. *Logistic Regression*

Logistic regression, a statistical method, is employed in music analysis to discern the factors influencing song virality [24]. With independent variables including 'beat', 'artist', and 'shareGroup', the dependent variable is 'virality'. Using Python, the data was processed by converting independent variables to categorical data and converting dependent variables to binary. After removing irrelevant columns, a design matrix was created using the patsy library, dividing the data set into dependent and independent variables for analysis.

After that, use the statsmodels library to create and fit a logistic regression model to data defined in terms of variables y and X. The sm.Logit function is used to create a logistic regression model, and the .fit() method to fit it to the data. This produces a result object (res) containing details about the adjusted model, such as coefficients for each independent variable, statistical significance, etc.

F. *Evaluation*

In the post-modeling phase, the evaluation process involves generating a model summary using the summary function. This summary includes essential evaluation metrics such as Df Residuals, Df Model, Pseudo R-squ, Log-Likelihood, LL-Null, LLR p-value, coefficient, standard error, t, P>|t|, upper bound, and lower bound [25]. These metrics collectively offer insights into model fit, the significance of independent variables, and the model's capacity to elucidate data variability.

IV. RESULT AND DISCUSSION

A. *Visualization*

The goal of using visualizations such as bar charts, pie charts, and correlation matrices is to understand patterns and relationships in data [26]. Bar and pie charts facilitate comparison of distributions of categorical variables, while correlation matrices reveal interactions between numerical variables. This helps researchers identify trends, gaps, and correlations, supporting accurate analysis of content virality on TikTok.

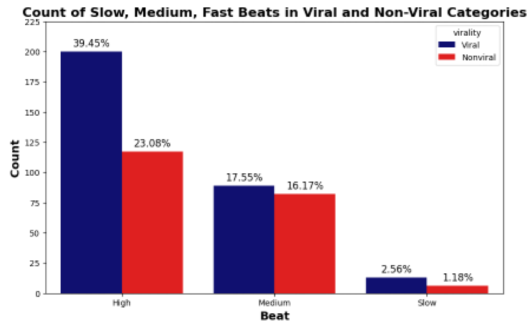

Fig 2. Percentage of Beats in Viral and NonViral Categories

From the visualization of the data observed in Figure 2, it can be seen that there is a correlation between the beat tempo of the video and the possibility of the video going viral. High tempo videos go viral more often than non-viral videos, with a percentage of 39.45% for viral videos and 23.08% for non-viral videos. Meanwhile, medium and slow paced videos rarely go viral, and slow paced videos have the lowest percentage in both categories. This suggests that high tempo may be a contributing factor to video virality, although other factors also play a role.

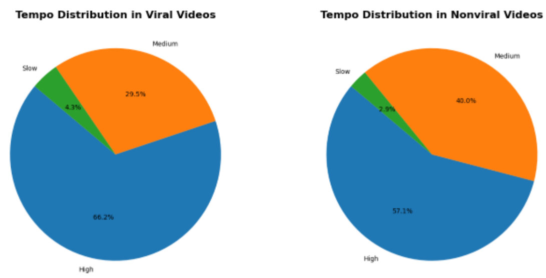

Fig 3. Tempo Distribution in Viral Videos dan Non-Viral Videos

The analysis of tempo distribution in TikTok videos, illustrated in Figure 3, underscores the significant influence of music on user engagement. The pie chart reveals a distinct difference in tempo distribution between viral and non-viral videos, with 66.2% of viral content featuring a high tempo. In contrast, non-viral videos exhibit a more balanced distribution between high (57.1%) and medium (40%) tempo. Slow tempo has minimal representation in both categories. This suggests a heightened potential for virality in videos with a high tempo. Practical implications for content creators include strategically choosing high-tempo songs to enhance engagement and increase the chances of content going viral. The energetic nature of up-tempo music may trigger spontaneous interactions, such as likes and shares, making it a crucial element in TikTok content strategy for maximizing virality and engagement.

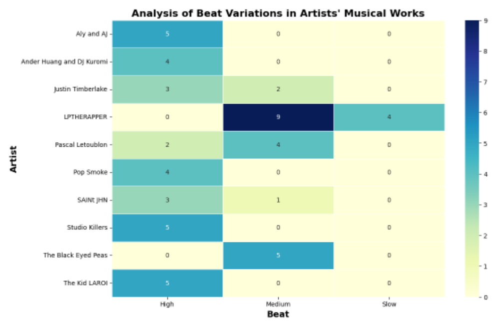

Fig 4. Analysis of Beat Variations in Artists' Musical Works

An exploration into potential variations in music tempo based on artists involved analyzing the top 10 artists with the highest music production. The assessment, depicted in Figure 4, showcased considerable diversity in tempo among these artists. LPTHERAPPER emerged prominently, contributing 13 music productions, with 9 featuring a medium beat tempo and 4 with a slow beat tempo. However, the majority of artists tended to maintain a consistent tempo across their musical works. Despite valuable insights, the analysis faced challenges due to limited song data for some artists, especially those with only one song in the dataset. This constraint hindered a comprehensive understanding of music tempo preferences.

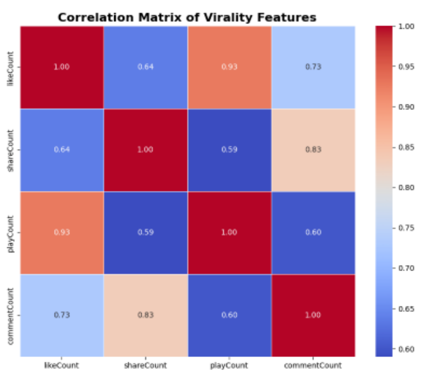

Fig 5. Correlation Matrix

The correlation analysis in Figure 5 underscores the pivotal role of "likes" in determining content virality, serving as a metric to gauge music's impact on audience response. Analyzing "likeCount" further allows researchers to categorize content into "Viral" or "Nonviral," making it the primary target variable. The observed correlations reveal a significant association between "likeCount" and other variables, particularly with values exceeding 0.5. While "shareCount" remains tolerable with a correlation of 0.64, the robust correlation of 0.93 between "playCount" and "likeCount" suggests a very strong relationship. Similarly, the correlation of 0.73 between "commentCount" and "likeCount" indicates a noteworthy association. The consideration of whether "playCount" and "commentCount" can serve as additional input variables for the analytical model is crucial.

B. Logistic Regression Summary Model

In this analysis, logistic regression was employed to investigate the factors influencing virality expectations, specifically artists, beats, or shares. The results of the analysis are presented below.

Logistic Regression Results						
Df Residuals	503					
Df Model	3					
Pseudo R-squ.	0.3123					
Log-Likelihood	-235.26					
LL-Null	-342.09					
LLR p-value	4.736e-46					
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.7279	0.244	-2.984	0.003	-1206	-0.250
artist	0.0008	0.001	0.750	0.453	-0.001	0.003
beat	-0.3214	0.205	-1.570	0.117	-0.723	0.080
shareGroup	2.4107	0.235	10.276	0.000	1.951	2.870

Fig 6. Logistic Regression Results

The logistic regression analysis results indicate that the variable most influencing virality in this dataset is 'shareGroup'. The regression coefficient for 'shareGroup' is 2.4107, and it's very low p-value (< 0.05) suggests that the relationship between 'shareGroup' and 'virality' is statistically significant. In other words, when 'shareGroup' increases, the likelihood of virality also significantly increases.

On the other hand, the variable 'beat' has a coefficient of -0.3214 with a p-value of approximately 0.117, indicating that the relationship between 'beat' and 'virality' is not statistically significant at the commonly accepted level of confidence ($\alpha = 0.05$). The variable 'artist' also does not have a significant impact on 'virality' with a coefficient of 0.0008 and a p-value around 0.453.

Furthermore, the pseudo R-squared value of 0.3123 suggests a moderate goodness of fit for the model, meaning that approximately 31.23% of the variability in 'virality' can be explained by the independent variables included in the model. Overall, these findings provide insight that the 'shareGroup' factor plays a significant role in increasing the likelihood of content going viral, while 'beat' and 'artist' do not show significant influence in this context.

V. CONCLUSION

This study explores the dynamics of music content virality on TikTok, focusing on beat, artist, and social shares. The analysis, utilizing the librosa library for tempo extraction and employing logistic regression, investigates the relationship between these factors and content virality, categorized by likes exceeding 5000.

As a result of the analysis, the variable 'shareGroup' significantly influences virality with a regression coefficient of 2.4107 and a p-value < 0.05, indicating that an increase in 'shareGroup' is positively associated with an increase in virality. Meanwhile, 'beat' is not significant (coefficient -0.3214, p-value 0.117), and 'artist' also does not have a significant impact (coefficient 0.0008, p-value 0.453) on virality. The pseudo R-squared value of 0.3123 indicates a moderately good fit for the model, explaining approximately 31.23% of the variability in virality through the included independent variables.

Looking ahead, future research could delve deeper into TikTok's social dynamics, unveiling intricate video-sharing mechanisms among users with similar music preferences. Such exploration may provide valuable insights for refining content strategies and optimizing sharing potential on the platform.

REFERENCES

[1] A. Bhandari and S. Bimo, “Why’s Everyone on TikTok Now? The Algorithmized Self and the Future of Self-Making on Social Media,” Social Media and Society, vol. 8, no. 1, 2022, doi: 10.1177/20563051221086241.

[2] A. Vizcaíno-Verdú and I. Aguaded, “#ThisIsMeChallenge and Music for Empowerment of Marginalized Groups on TikTok,” Media Commun, vol. 10, no. 1, 2022, doi: 10.17645/mac.v10i1.4715.

[3] R. Panda, R. Malheiro, and R. P. Paiva, “Audio Features for Music Emotion Recognition: A Survey,” IEEE Trans Affect Comput, vol. 14, no. 1, pp. 68–88, Jan. 2023, doi: 10.1109/TAFFC.2020.3032373.

[4] Y. Xia and F. Xu, "Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm," Mathematical Problems in Engineering, 2022.

[5] M. Y. Rodriguez and H. Storer, "A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data," Journal of Technology in Human Services, vol. 38, no. 1, pp. 54-86, 2020.

[6] O. Sadler, “Defiant Amplification or Decontextualized Commercialization? Protest Music, TikTok, and Social Movements,” Social Media and Society, vol. 8, no. 2, 2022, doi: 10.1177/20563051221094769.

[7] Y. Wang, “Contagion effect of music beats on audient emotions,” Revista Argentina de Clinica Psicologica, vol. 29, no. 1, 2020, doi: 10.24205/03276716.2020.155.

[8] H. Yang, “Research on Music Advertising in TikTok under a Systematic View,” Advances in Education, Humanities and Social Science Research, vol. 4, no. 1, 2023, doi: 10.56028/aehssr.4.1.381.2023.

[9] T. Cook, A. R. K. Roy, and K. M. Welker, “Music as an emotion regulation strategy: An examination of genres of music and their roles in emotion regulation,” Psychol Music, vol. 47, no. 1, 2019, doi: 10.1177/0305735617734627.

[10] B. Radovanović, “TikTok and Sound: Changing the ways of Creating, Promoting, Distributing and Listening to Music,” INSAM Journal of Contemporary Music, Art and Technology, no. 9, 2022, doi: 10.51191/issn.2637-1898.2022.5.9.51.

[11] T. Shaikh and A. Jadhav, “Music Genre Classification Using Neural Network,” ITM Web of Conferences, vol. 44, 2022, doi: 10.1051/itmconf/20224403016.

[12] A. Vizcaíno-Verdú, P. De-Casas-moreno, and S. Tirocchi, “Online prosumer convergence: Listening, creating and sharing music on YouTube and TikTok,” Communication and Society, vol. 36, no. 1, 2023, doi: 10.15581/003.36.1.151-166.

[13] K. Choi, G. Fazekas, K. Cho, and M. Sandler, “The Effects of Noisy Labels on Deep Convolutional Neural Networks for Music Tagging,” IEEE Trans Emerg Top Comput Intell, vol. 2, no. 2, 2018, doi: 10.1109/TETCI.2017.2771298.

[14] I. L. Simarmata and I. W. Supriana, “Music Genre Classification Using Random Forest Model,” JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), vol. 12, no. 1, 2023, doi: 10.24843/jlk.2023.v12.i01.p10.

[15] E. van de Ven, "TikTok Trending Videos," Kaggle, December 2020. [Online]. Available: https://www.kaggle.com/datasets/erikvdven/tiktok-trending-december-2020?select=videos.

[16] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, “Exploratory data analysis using python,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 12, 2019, doi: 10.35940/ijitee.L3591.1081219.

[17] B. Rafaely *et al.*, “Spatial Audio Signal Processing for binaural reproduction of recorded acoustic scenes – review and Challenges,” *Acta Acustica*, vol. 6, p. 47, 2022.

[18] “Top Popular Python Libraries in Research,” Journal of Robotics and Automation Research, vol. 3, no. 2, 2022, doi: 10.33140/jrar.03.02.02.

[19] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, “Data Preprocessing,” in International Series in Operations Research and Management Science, vol. 334, 2022. doi: 10.1007/978-3-031-16990-8_4.

[20] S. Sakthi Vel, "Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 879-884, doi: 10.1109/ICAIS50930.2021.9395924.

[21] P. A. Abdulla, M. F. Atig, Y. F. Chen, B. P. Diep, J. Dolby, P. Janků, dan W. C. Wu, "Efficient handling of string-number conversion," in Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation, June 2020, pp. 943-957.

[22] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, “Exploratory data analysis using python,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727–4735, Oct. 2019, doi: 10.35940/ijitee.L3591.1081219.

[23] IBM, “What is logistic regression?,” IBM.com.

[24] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, “Logistic Regression,” in *International Series in Operations Research and Management Science*, vol. 334, 2022. doi: 10.1007/978-3-031-16990-8_7.

[25] A. A. T. Fernandes, D. B. F. Filho, E. C. da Rocha, and W. da Silva Nascimento, “Read this paper if you want to learn logistic regression,” *Revista de Sociologia e Politica*, vol. 28, no. 74, pp. 1/1-19/19, 2020, doi: 10.1590/1678-987320287406EN.

[26] Lu, M., Lanir, J., Wang, C., Yao, Y., Zhang, W., Deussen, O., & Huang, H. (2022). Modeling just noticeable differences in charts. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 718–726. https://doi.org/10.1109/tvcg.2021.3114874.