INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Predicting Music Popularity Using Spotify and YouTube Features

Yap Kah Yee[1]*, **Mafas Raheem**[1]

1 School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia

OPEN ACCESS

Received: 13-12-2021

Accepted: 23-08-2022

Published: 21-09-2022

Citation: Yee YK, Raheem M (2022) Predicting Music Popularity Using Spotify and YouTube Features. Indian Journal of Science and Technology 15(36): 1786-1799. https://doi.org/10.17485/IJST/v15i36.2332

* **Corresponding author**.

rmafas@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To examine whether the integration of Social Media features from YouTube videos and Spotify audio features can effectively predict music popularity. **Methods:** A dataset is constructed by collecting newly released tracks from May to August 2021. Audio features are acquired from Spotify while social media features are obtained from the official videos on YouTube. Music popularity is defined using five metrics derived from the Spotify Top 200 daily chart performance to measure diverse aspects of the songs' success (Length, Max, Sum, Mean, and Debut). The predicted popularity has three target variables, ranging from Low, Medium to High popularity. During model implementation, four machine learning models were trained on the dataset in two different stages such as purely audio features and both audio and social media features respectively. **Findings:** At the second stage, random forest outperformed the other three models with the best results for the four-evaluation metrics. In detail, the model generated accuracy of 79.6%, macro-precision of 74.5%, macro-recall of 73.2%, and macro F1-scores of 73.1% on average across the five-popularity metrics used. Moreover, the results from both experimental stages showed that the incorporation of social media variables significantly increased the model performances relative to the use of audio features only, with the margins of improvement ranging from 10% to 60%. This demonstrates that YouTube-based social media features are beneficial for the use of industry practitioners to identify potentially popular hits. **Novelty:** This research appears to be the first study to date in the Hit Song Science domain that utilizes Social Media data from YouTube for the prediction of hit songs. Furthermore, it promotes the prediction of potential hits by using audio features and social media data jointly.

Keywords: hit song science; machine learning; audio features; social media features; Spotify; YouTube

1 Introduction

What makes a song a hit? Researchers have conducted studies to develop a better understanding and investigate the relationship between the intrinsic quality of songs and their popularity. This is known as Hit Song Science (HSS), a subfield under the

Music Information Retrieval domain, an interdisciplinary science on analyzing and retrieving information from music. The reason for the sustained interest in HSS studies boils down to its practical implications for the industry stakeholders. If there are indeed certain characteristics that can guarantee the success of the song pre-release, record labels and artists are better positioned to mitigate massive monetary costs that are usually involved.

Nevertheless, the music industry has witnessed a significant change in terms of how people consume music in recent years. The consumption of physical compact discs has dramatically declined and replaced by music streaming online as a variety of music streaming platforms emerged with the likes of Spotify, YouTube Music, Apple Music, and others. For instance, YouTube disclosed that they have paid out US$4 billion to the likes of artists, songwriters, and rights-holders in 2020 [(1)]. The sharing of music videos on YouTube has also catapulted songs to huge success and fame, especially for lesser-known artists. An exemplary example is 'Despacito' by Spanish artists Luis Fonsi and Daddy Yankee, which became the most viewed music video on YouTube with over seven billion views in 2020 [(2)].

In light of these changes in the industry, the more recent HSS studies have yet to incorporate these trends or developments and utilized more up-to-date features in their predictions. Although social media data have been explored in prior studies, the use of YouTube features specifically is not observed yet. On the other hand, YouTube-based social media features are more commonly observed in other cultural markets such as the movie and fashion industry and have proven to be effective for the prediction of real-world outcomes. The study by Ahmad et al [(3)] demonstrated that variables such as the number of views, comments, likes, and dislikes for movie trailers are effective for box-office revenue prediction.

Hence, this research seeks to develop a more profound understanding of the benefits of utilizing social media data to predict potential hits. Therefore, industry stakeholders can prioritize more effective measurement of opinions and post-release feedback, which they can swiftly respond. Indeed, Watts & Hasker [(4)] acknowledged the effect of social influence and the authors opined that record label executives should take this factor into account while devising their marketing strategies. For instance, they recommended focusing on real-time response tracking and building more flexibility into their allocation of marketing resources according to the consumers' reactions or demand.

Hence, the main contribution of this research is it appears to be the first study to date, in the HSS domain, that utilizes social media data from YouTube for the prediction of hit songs. While previous studies have only looked at hit song predictions using either audio features or social media features independently, this research promotes the prediction of potential hits from a more holistic perspective by using audio features and social media data jointly.

2 Literature Review

Various studies in HSS have attempted to identify features that can explain songs' popularity, primarily via internal and external perspectives. Internal perspective refers to the innate characteristics of songs such as audio features and lyrics whereas external aspects describe the musical ecosystem, for instance, social or market data. To date, the majority of the studies have focused on using mostly audio features [(5–9)]. These studies demonstrated that acoustic or audio features are influential in predicting songs' success with good accuracy levels. A small number of early studies used lyrics [(10)] or social media data [(11,12)] to predict the popularity of songs. In particular, the role of social influence is more evident in predicting a song's popularity, in the form of early chart performance or social media content. Studies that tapped into the use of social media content, namely Last.FM listener behaviour and tweets relevant to the users' listening activities from Twitter, for hit songs predictions, have demonstrated that they boosted the prediction accuracies [(11,12)]. Nevertheless, both audio and social media aspects were not examined together in these studies, not to mention that the social media platforms used are not essentially music streaming services. Twitter, a microblogging platform, allows users to share the music that they are listening to via tweets. Last.FM mainly lets users track their music playback, although they have recently allowed playing music on its site through Spotify. Thus, these platforms would not be able to provide an accurate and comprehensive picture of the users' music streaming activities. On the contrary, social media variables such as YouTube-based variables are adopted more frequently for the prediction of box office performances in the movie industry and are found to be influential in movie revenue prediction [(3)]. This suggested that a similar approach can be extended to the music industry as well to forecast real-world outcomes.

3 Methodology

Figure 1 depicts a flow chart of the research methodology. Relevant data were gathered from multiple sources, namely Spotify and YouTube. The tracks along with their audio and social media features were combined into an Excel file. Daily chart performances from Spotify were also downloaded during the period and various performance metrics were computed. The first dataset containing the tracks and their corresponding features were then matched against the second dataset containing the chart performances. As a result, the final dataset is formed.

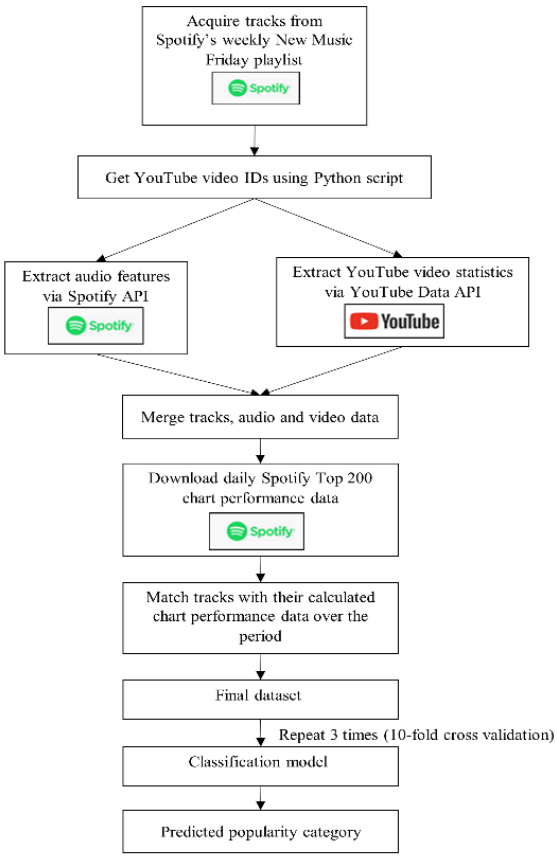

Fig 1. Flow chart of the research methodology

3.1 Data collection

The research is conducted using newly released tracks from 14 May to 20 August 2021, which were compiled based on the New Music Friday playlist, a weekly collection of 100 new tracks curated by Spotify. The playlist serves as a representative sample of tracks that covers a wide range of genres and artists, including some lesser-known, independent artists. Audio and social media data were acquired using automated scripts via Spotify and YouTube Data APIs. Only newly released songs were considered due to the constraint that historical YouTube videos statistics for social media features could not be retrieved using their API except for the channel owner or content manager. Eventually, a total of 1,432 unique tracks are left in the dataset.

3.2 Dependent variables

The main target variables for this study are the songs' popularity metrics based on the daily Spotify Top 200 chart performances. While the majority of existing studies preferred to use Billboard chart performance as their dependent variable[13,14], Spotify Top 200, which displays the platform's top 200 songs with the highest daily number of streams, might be the more suitable choice given that it is one of the most relevant music charts at present. Daily top 200 charts are parsed from the Spotify website from 15 May to 1 September 2021, containing the track's title, rank, artist, number of streams and dates.

So far, most HSS studies predict on a song's existence to the music chart or not. However, this research decided to use the approach in Lee & Lee[15] to develop deeper insights by measuring diverse aspects of popularity. These insights gained can be used in better planning of the marketing strategy or allocation of budget and other resources. When only a single aspect of music popularity is considered, valuable information is lost. For example, a viral song may stay in the chart for only a week before it quickly loses its popularity and drops out of the chart after that. Multiple metrics can be extracted using rankings of songs in a music chart over time to describe different aspects of a song's popularity. These popularity metrics are measured

using the rank score, which is defined as in (1), where Max_rank is the lowest possible rank of the chart and rank (i) is the rank of the song.

$$Rank\_score\ (i) = Max\_rank - rank\ (i) + 1$$

Table 1 displays the different popularity metrics developed by the authors[15] and measured using the computed rank scores where Length, Max, Sum, Mean and Debut are used for this research. Similar to the approach used by the authors, hit song prediction is modelled as a classification task in this research. Binary classification of each of the popularity metrics was performed with the median rank score as the designated boundary separating the two classes. However, certain adjustments were made in this research, given the tracks that could not make it to the chart and thus, were assigned a value of 0 concerning their performance metrics.

In this research, the prediction of music popularity was framed as a multi-class classification problem where the tracks have either Low, Medium, or High popularity concerning Length, Max, Sum, Mean and Debut. For model building purposes, the target classes were assigned with three levels such as 1 for Low popularity, 2 for Medium popularity and 3 for High popularity.

Table 1. Popularity Metrics

Metrics	Descriptions
Debut	Defined as the rank score of a song when the song first appears in the chart. Indicates the initial popularity of the song.
Length	Defined as the time during which a song appears on a chart. Measures how long has a song been popular.
Kurtosis	Describes the pattern of growing and declining popularity together with skewness. Higher values indicate faster popularity of a song's growth.
Max	Defined as the maximum rank score of a song during the entire period. Indicates the peak popularity of the song.
Mean	Defined as the average rank score of a song over the entire period during which the song appears in the chart.
Skewness	Describes the dynamic pattern that a song gains and loses popularity. A positive value indicates that the song becomes popular quickly, reaches its peak popularity, and decays slowly. A negative value indicates that the song becomes popular slowly and loses its popularity quickly.
Standard Deviation	The standard deviation of the rank score of a song over the entire period during which the song appears in the chart. Represents the change in popularity of a song over time.
Sum	Defined as the sum of the rank scores over time. Describes the overall popularity of a song during the whole period.

3.3 Independent variables

A set of audio features that is easily interpretable by industry stakeholders was acquired from Spotify API using the New Music Friday playlist ID weekly. The majority of them are numerical variables, except mode and key, which are found as categorical. These same features were also used in existing studies and have shown good results while showcasing their significance[13,14,16]. The lists of features together with their descriptions[17] are tabulated in Table 2 .

Most of the weekly new releases in Spotify can be located and collected on YouTube. Based on the tracks acquired earlier using Spotify New Music Friday playlist, the title of the tracks was used as keywords to retrieve the most relevant YouTube video IDs. Hence, four relevant YouTube-based metrics were identified for this research, comprising view metrics and engagement metrics. View metrics include several video views, which is typically the central gauge of popularity in YouTube. Engagement metrics are some comments, likes and dislikes. These metrics got downloaded on the 7^{th} day after the videos were released, with the help of a scheduler setup in API connector for Google Sheets.

On top of the four variables, additional engagement metrics were also used to compare videos with varying absolute amounts of views, comments, or votes, as proposed by Liikkanen[18]. These metrics are the number of votes and comments per thousand views to measure the frequencies of voting and commenting as well as the dislike proportion to represent the share of dislikes. In a later study, these metrics were also adopted to better analyze and compare engagement statistics for music videos on YouTube[19].

3.4 Predictive machine learning algorithms

The predictive machine learning algorithms were chosen from the existing literature such as random forest, logistic regression, support vector machine and neural networks and consequently to advocate the best performing model in this regard.

Random forest is a robust classifier made up of a collection of tree-structured classifiers, which tends to produce more favourable results than individual classifiers as a large number of trees votes for the most popular class[20]. In several studies,

Table 2. Spotify Audio Features

Features	Data Types	Descriptions
Dance-ability	Float	Describes how suitable a track is for dancing based on a combination of musical elements. A value of 0.0 is least danceable and 1.0 is the most danceable.
Energy	Float	A measure from 0.0 to 1.0 represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
Key	Inte-ger	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.
Loudness	Float	The overall loudness of a track is decibels. Values typical range between -60 and 0 dB.
Mode	Inte-ger	Indicates if the track is major or minor. Major is represented by 1 and minor is 0.
Speechi-ness	Float	Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values below 0.33 most likely represent music and other non-speech-like tracks.
Acoustic-ness	Float	Describes whether the track is acoustic-based on a confidence measure from 0.0 to 1.0. 1.0 represents high confidence the track is acoustic.
Instru-mental-ness	Float	Predicts whether the track contains no vocals. Higher instrumentalness values represent a higher probability that the track contains no vocal content.
Liveness	Float	Detects the presence of an audience in the recording. Higher liveness values represent a higher probability that the track was performed live.
Valence	Float	A measure from 0.0 to 1.0 describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive, while tracks with low valence sound more negative.
Tempo	Float	The overall estimated tempo of a track in beats per minute.

random forest performed relatively well compared to the other models with accuracies above 80%[1,14]. The model is generally applicable to a wide range of prediction problems and capable of dealing with smaller sample sizes and high-dimensional feature spaces[21].

Logistic regression is a popular algorithm to use when the outcome variable is dichotomous or binary, although it can also be extended for multi-class output. Essentially, logistic regression uses the sigmoid activation function so that the class probabilities are between 0 and 1. Logistic regression also comes across as another popular model of choice for hit song predictions. Even though it does not perform as reasonably well as random forest, the algorithm is adequately cited by some researchers in their studies[7,13].

Support vector machine is also often selected by researchers for hit song predictions studies[15,16]. Support vector machine has been well regarded as one of the most robust machine learning techniques. It works by selecting the most optimal hyperplane, which produces the maximum margin that separates the two classes.

Neural networks are less frequently featured in earlier studies. Nonetheless, as seen in more recent research[4,5], neural network architecture achieved satisfactory results. A neural network is particularly effective for modelling complex, non-linear functions.

3.5 Experimental setup and evaluation

Two experiments were carried out to assess the significance of the features used for predictions. In the first experiment, the machine learning algorithms were trained using only the audio features. The second experiment evaluated the performance improvement by including the additional social media features.

The entire dataset was split into training and testing sets following an 80:20 split. 10-fold cross-validation was applied with three repeats to validate the model performance. At the same time, hyper-parameter tuning was also carried out to identify the most optimal hyper-parameters for each of the algorithms, via GridSearchCV or RandomizedSearchCV functions.

Evaluation measures such as accuracy, recall, precision, and F1-scores were used to evaluate the performance of the models. To be precise, macro-recall, macro-precision and macro F1-scores were used given that the prediction problem is structured as multi-class classification. Generally, there are two different methods to compute macro F1-scores. The formula used here is to average the computed F1-scores for each class via arithmetic mean. This computation yields more robust results when the

dataset has an imbalanced issue, relative to the other method which produces high evaluation scores that are misleading[22]. Although accuracy is computed as one of the evaluation measures, the other three measures are deemed as more appropriate in the hit song prediction. This is on the assumption that the cost of incorrectly identifying a song is considerably high in terms of resources spent.

4 Implementation

4.1 Data preparation

Few data pre-processing steps were implemented to prepare the data. Missing data handling was not required as no missing values were detected in this dataset. Further, the independent variables were normalized before training the models since the features had significant differences in their scales, especially with the social media variables ranging up to millions.

As mentioned earlier, three additional variables were also incorporated on top of the four YouTube-based features such as number of votes per thousand views, number of comments per thousand views and dislike proportion. Equations (2), (3) and (4) illustrate the computations of these variables.

$$Number\ of\ votes\ per\ 1000\ views = \frac{Number\ of\ votes * 1000}{Number\ of\ views}$$

$$Number\ of\ comments\ per\ 1000\ views = \frac{Number\ of\ comments * 1000}{Number\ of\ views}$$

$$Dislike\ proportion = \frac{Number\ of\ dislikes}{Number\ of\ likes + Number\ of\ dislikes}$$

4.2 Data analysis

An exploratory data analysis was performed to better understand the predictor and the dependent variables. Although each of the dependent variables was split into three categories of popularity for prediction, to ease the comparison for data analysis purposes, the tracks were split into two groups such as those that make it to the chart and those that do not.

On a broader level, the nine audio features provided by Spotify can be grouped into three general categories such as mood, properties, and context.

- Spotify describes the mood of a song by calculating its danceability, valence, energy, and tempo. The tracks included in the dataset are relatively more danceable and energetic with mean values hovering within the range of 0.6 to 0.7. In terms of valence or positiveness in simpler terms, the tracks are more or less neutral. Tempo-wise, the average is about 122 beats per minute. Interestingly, according to an article by BBC music reporter, the tempo for the top 20 best-selling songs each year has been on an increasing trend and coincidentally, the average tempo of 2020's top bestsellers is a 'pulse-quickening 122 beats per minute'[16,23].
- On the other hand, the properties of a track are defined by loudness, speechiness and instrumentalness. The tracks on average have a loudness value of -6.7 decibels and are neither speech-like nor instrumental with low mean values of 0.12 and 0.03 respectively.
- The context of a song is represented by its liveness and acousticness values. Judging by their low mean values, the tracks are less likely to be acoustic as well as performed live.

The remaining two categorical audio features were also analysed. About 64% of the tracks are in major keys. The tracks are also fairly distributed across 12 keys with the majority of them in keys 0 and 1. As for the social media features, the descriptive statistics in Table 3 shows that the four variables have positively skewed distributions with significantly greater means than medians. For the newly added social media features, the distributions are also positively skewed but less extreme compared to the original variables.

The predictor variables were compared with the tracks categorized as non-hits and hits using boxplots for numerical variables and stacked bar charts for categorical variables. Results showed that there is no discernible pattern that separates hits from non-hits. In contrast, there is a noticeable trend in the boxplots for social media variables, which are log-transformed due to their highly skewed nature. Based on Figure 2 , the median view counts of hit songs are significantly higher compared to non-hits.

Table 3. Descriptive statistics of social media features

Social Media Features	Min	Median	Mean	Max
Number of video views	41	99,688	1,615,387	244,704,749
Number of likes	1	4,126	80,578	15,699,273
Number of dislikes	0	61	1,455	179,519
Number of comments	0	260	10,925	6,638,156
Likes per 1000 views	0	47	52	282
Comments per 1000 views	0	3	4	68
Dislike proportion (%)	0	1	2	50

The same holds for comment, like and even dislike counts as displayed in Figure 2 and Figure 3. This is within expectations. Logically, videos that attracted a higher absolute number of views should also receive more comments and votes, both positive and negative. Interestingly, the trend for the added social media variables is muted with no distinct differences between hits and non-hits (Figures 4 and 5). To summarize, these visualizations provided a better gauge of the potentially influential factors for predicting music popularity during model building.

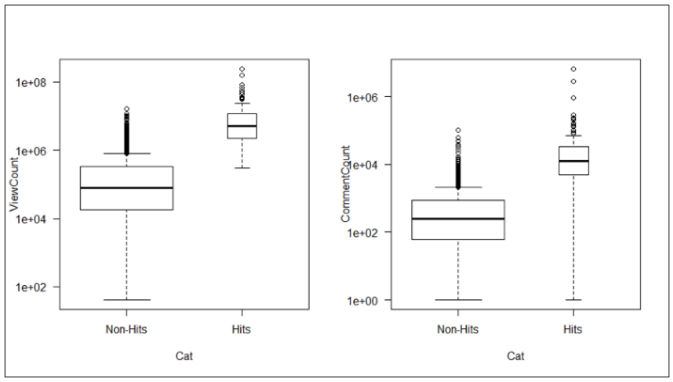

Fig 2. Boxplots of number of views and comments against hits and non-hits

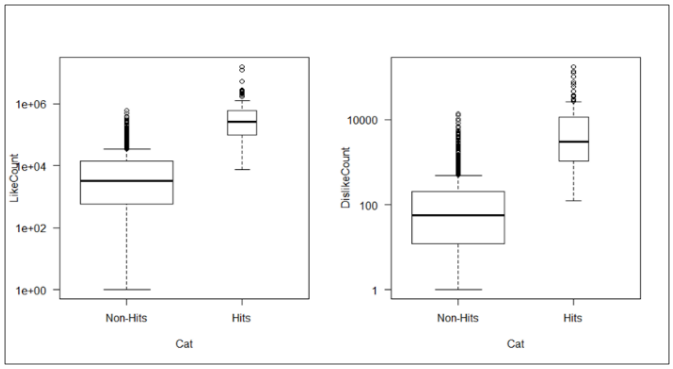

Fig 3. Boxplots of number of likes and dislikes against hits and non-hits

Correlation heatmap was used to check for existence of strong correlations between the numerical independent variables (Figure 6). As anticipated, view counts are highly correlated with comment, likes and dislike counts. The latter variables were removed from the dataset. The frequencies of commenting and voting were better measured by the new social media variables. Even though certain pairs of variables such as energy and loudness as well as comments and votes per thousand views displayed moderate strong positive linear relationship and were retained in the dataset as their degree of correlations are still acceptable.

Concerning the dependent variables, the tracks were analysed based on their chart performances. Table 4 displays the popularity distribution of tracks, which were derived by summing up their target class output (values from 1 to 3) across the

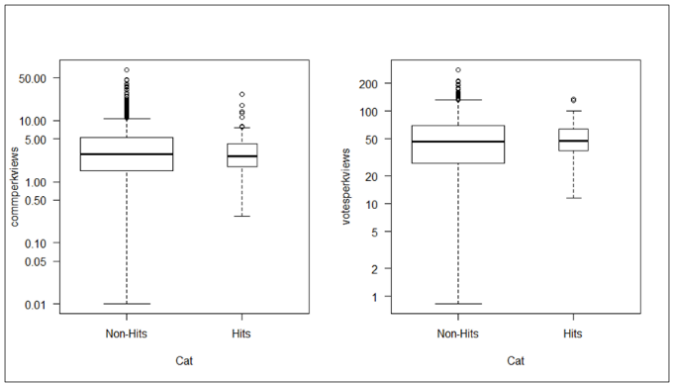

Fig 4. Boxplots of comments per thousand views and votes per thousand views against hits and non-hits

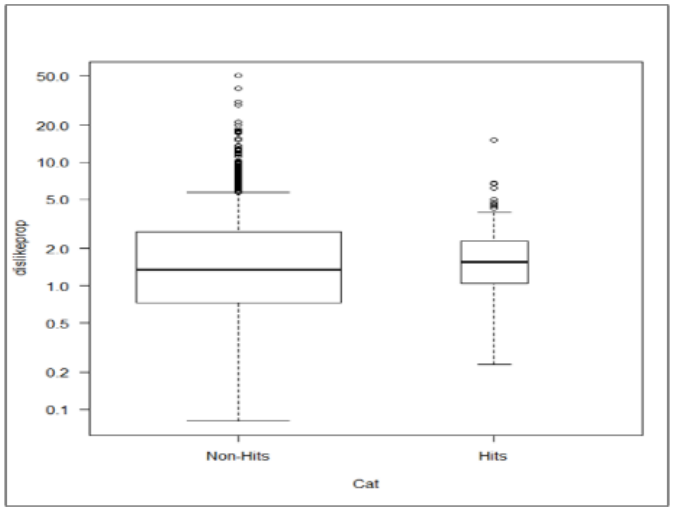

Fig 5. Boxplots of dislike proportion against hits and non-hits

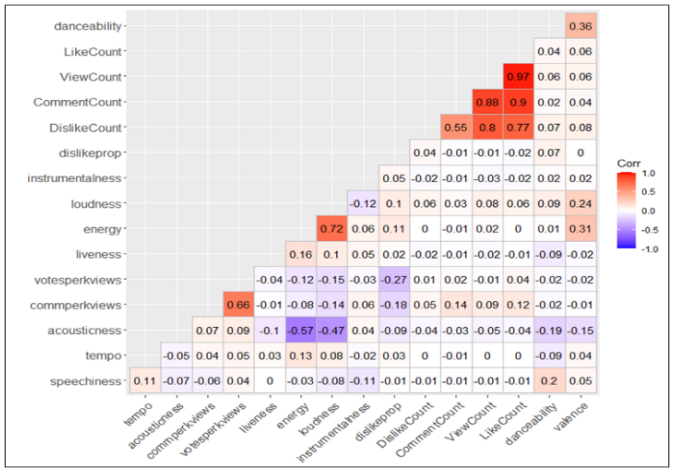

Fig 6. Correlation heatmap of numerical independent variables

five dependent variables. Hence, a track that did not appear in the chart got a total popularity score of 5 (Low popularity with the class output of 1) whereas a track that made it to the chart could get a score from 10 to 15. Out of the 1,432 tracks, only 127 tracks were able to make it to the chart and among these songs, only 34 tracks have achieved High popularity for all five popularity aspects.

Table 4. Popularity distribution of tracks

Total Popularity Score	Number of Tracks	Percentage of Tracks (%)
5	1305	91.13%
10	35	2.44%
11	15	1.05%
12	16	1.12%
13	10	0.70%
14	17	1.19%
15	34	2.37%

In addition, the individual popularity metrics were examined further. Concerning the target variable Length, approximately 91% of these tracks did not appear on the Spotify Top 200 chart throughout the observed period. On the other hand, two tracks managed to stay within the chart for 110 days, which is also the total duration of the data collection period. In terms of Max, only two songs became the number one hits on the Spotify chart. The histograms for Sum and Mean also displayed similar positively skewed distributions, with only a small percentage of tracks achieving high popularity. For Debut, the highest position that the songs have debuted is second on the chart and only two songs in total have accomplished that. These findings also suggested that the chances of producing a chart-topping hit song are very slim, which sheds light on the rationale behind HSS.

At this juncture, it should be apparent that this dataset is highly imbalanced with only a small fraction of hit songs. To avoid the issue a random under-sampling method was used to eliminate observations from the majority class, which are those with Low popularity. After undersampling, the dataset consisted of 127 tracks under Low popularity, 64 under Medium popularity and 63 under High popularity for all five dependent variables. Although there was still a slight imbalance issue for the Low popularity class, the distribution was more balanced in this scenario such that the proportion of non-hits (Low popularity) and hits (Medium and High popularity combined) were equal. This was also in consideration that the dataset would be too small if the number of tracks in the Low popularity class are reduced to be equivalent to that of Medium or High popularity. Finally, the dataset left with 254 tracks for model building.

4.3 Model implementation

As mentioned, two experimental stages were carried out sequentially. The first stage involved model building using only audio features while the second stage included both audio and social media features.

Firstly, the random forest model was built using the selected features while having the hyper-parameters such as ntree and mtry tuned. Secondly, the logistic regression model was built with hyperparameter tuning by taking the multi-class output as target ordered from Low to High popularity. Thirdly, the support vector machine model was built using three different kernel choices such as linear, radial, and polynomial. The appropriate parameters were also tuned for each of these kernels to derive the best-performing ones. Finally, the neural network model was built with hyperparameter tuning for the number of units in the hidden layer, decay, and the regularization parameter that mitigates overfitting. Table 5 illustrates the optimal parameters used after tuning for each of the models.

5 Results and Discussion

5.1 Model results

Tables 6, 7, 8 and 9 summarize the comparison of performance evaluation metrics for the four models. The focus is more on macro-precision, macro-recall, and macro F1-scores, given that they are more relevant for the problem domain due to an imbalanced issue associated with the Low popularity category in the dataset. The best performing model at the two experimental stages is highlighted in bold for each of the tables. Nevertheless, the macro-precision and macro F1-scores must be interpreted with caution. The precision scores for some of the classes (in particular Medium and High popularity categories) were undefined with no positive cases in the confusion matrix. For easier handling, they were treated as 0 and macro-precision values were derived by averaging across three classes. The same treatment was applied for macro F1-scores during scenarios when recall

Table 5. Optimal parameters for random forest post-tuning

Target Variables	Features Used	Random Forest		Ordinal Regression	Support Vector Machine			Neural Network	
		ntree	mtry	Method	Kernel	Gamma	Cost	Size	Decay
Length	Audio only	500	1	cloglog	Radial	0.4	10	1	0.1
Max	Audio only	500	1	loglog	Radial	0.4	1	1	0.1
Sum	Audio only	500	1	logistic	Radial	0.4	10	3	0.01
Mean	Audio only	500	1	probit	Radial	0.4	1	1	0.1
Debut	Audio only	500	1	cloglog	Radial	0.4	10	3	0.0
Length	Audio and social media	2000	10	cauchit	Linear	-	1000	1	0.01
Max	Audio and social media	2000	5	cauchit	Linear	-	10	1	0.01
Sum	Audio and social media	2000	10	cauchit	Linear	-	1000	1	0.01
Mean	Audio and social media	500	7	cauchit	Linear	-	100	1	0.01
Debut	Audio and social media	2000	9	cauchit	Linear	-	10	1	0.01

and precision were both 0 for certain classes. The evaluation metrics that were subjected to this issue got marked using asterisks (*).

Table 6. Comparisons of Accuracy for The Four Models

Target Variables	Features Used	Random Forest	Ordinal Regression	Support Vector Machine	Neural Network
Length	Audio only	50.0%	44.0%	**52.0%**	36.0%
Max	Audio only	50.0%	40.0%	**50.0%**	28.0%
Sum	Audio only	50.0%	46.0%	**52.0%**	36.0%
Mean	Audio only	50.0%	40.0%	**52.0%**	36.0%
Debut	Audio only	50.0%	44.0%	**54.0%**	30.0%
Length	Audio and social media	**82.0%**	70.0%	72.0%	36.0%
Max	Audio and social media	**76.0%**	64.0%	68.0%	76.0%
Sum	Audio and social media	**84.0%**	70.0%	72.0%	70.0%
Mean	Audio and social media	**80.0%**	68.0%	78.0%	76.0%
Debut	Audio and social media	**76.0%**	60.0%	66.0%	70.0%

Table 7. Comparisons of macro-precision for the four models

Target Variables	Features Used	Random Forest	Ordinal Regression	Support Vector Machine	Neural Network
Length	Audio only	16.7% *	26.9% *	**50.4%**	22.8% *
Max	Audio only	16.7% *	23.8%	**50.4%**	14.6% *
Sum	Audio only	16.7% *	28.4% *	**50.4%**	31.4%
Mean	Audio only	16.7% *	22.0% *	**50.7%**	22.1% *
Debut	Audio only	16.7% *	25.8% *	**72.9%**	27.9%
Length	Audio and social media	**78.6%**	65.9%	64.8%	19.9% *
Max	Audio and social media	**69.5%**	64.6%	64.8%	66.2%
Sum	Audio and social media	**81.7%**	68.9%	68.4%	62.5%
Mean	Audio and social media	**74.3%**	68.6%	75.3%	72.4%
Debut	Audio and social media	**68.2%**	55.4%	60.3%	57.8%

Table 8. Comparisons of macro-recall for the four models

Target Variables	Features Used	Random Forest	Ordinal Regression	Support Vector Machine	Neural Network
Length	Audio only	33.3%	33.5%	**36.8%**	29.2%
Max	Audio only	33.3%	27.2%	**33.9%**	18.7%
Sum	Audio only	33.3%	34.4%	**37.1%**	30.6%
Mean	Audio only	33.3%	28.4%	**35.6%**	27.6%
Debut	Audio only	33.3%	32.0%	**40.9%**	27.6%
Length	Audio and social media	**77.0%**	64.2%	63.2%	27.4%
Max	Audio and social media	**69.6%**	60.5%	63.7%	64.3%
Sum	Audio and social media	**79.6%**	66.6%	67.7%	60.4%
Mean	Audio and social media	**73.1%**	65.3%	74.9%	66.2%
Debut	Audio and social media	**66.4%**	54.2%	60.2%	58.4%

Table 9. Comparisons of macro f1-scores for the four models

Target Variables	Features Used	Random Forest	Ordinal Regression	Support Vector Machine	Neural Network
Length	Audio only	22.2% *	29.3% *	**30.6%** *	25.6% *
Max	Audio only	22.2% *	22.2% *	**25.7%** *	16.4% *
Sum	Audio only	22.2% *	29.4% *	**31.1%** *	30.5%
Mean	Audio only	22.2% *	23.7% *	**27.0%** *	24.4% *
Debut	Audio only	22.2% *	27.5% *	**37.0%**	27.6%
Length	Audio and social media	**77.7%**	64.3%	63.4%	23.0% *
Max	Audio and social media	**67.8%**	58.0%	61.4%	64.8%
Sum	Audio and social media	**80.2%**	67.1%	67.9%	55.8%
Mean	Audio and social media	**73.6%**	65.5%	75.1%	64.7%
Debut	Audio and social media	**66.3%**	54.2%	59.9%	55.6%

Out of the four models, the support vector machine consistently outperformed the others for solely audio features, across all four-evaluation metrics. However, random forest appeared to be the best performing model for the combination of audio and social media features. The overall results were poor for audio features only, thus support vector machine obtained 52.0% accuracy and 30.3% macro F1-scores on average across the five target variables. A closer examination of the confusion matrices revealed that most of the time, the models predicted almost all of the tracks in the Low popularity category. This suggests that the models are not capable of distinguishing hits from non-hits using purely audio features.

Nonetheless, there is a marked increase in prediction accuracy and the other evaluation metrics after incorporating social media features. In general, substantial improvement was observed for all of the four models, although the margins vary from 10% to 60% for the four metrics. The only exception is for the prediction of Length for the neural network, which did not observe any improvement in accuracy and performed worse in the other evaluation metrics at the second stage which contrasted sharply with the neural network results for the other four target variables. However, the random forest generated on average 79.6% accuracy and 73.1% macro F1-scores after the addition of social media variables, as compared to 50.0% accuracy and 22.2% macro F1-scores before. The improvements were noted from more accurate predictions of tracks under the Medium and High popularity category.

5.2 Discussion and Implications

The support vector machine was selected as the best model when only audio features were used. With the addition of social media features, the random forest turned out to be the best performing model. This seems consistent with existing studies which compared different machine learning models [3,16].

The experiments with solely audio features generated rather poor results, which are not better than random which raises the question of the effectiveness of HSS. The potential reason that accounted for the unsatisfactory performance of the models could be the size of the dataset. The results are similar to those obtained in one of the studies surveyed with the lowest accuracy values between 50% to 52% [24], which was conducted with a relatively small dataset containing 647 songs in total. In contrast, existing studies that utilized sufficiently large datasets with thousands of tracks were able to return considerably good accuracy,

precision and recall scores [3,16]. This indicates that adequate sample sizes should be able to capture the nuances that distinguish between hits and non-hits better. Regardless, it is perhaps not fair to benchmark the results against existing studies given that the dataset and features used are different.

Despite the poor results during the first stage, the models eventually returned satisfactory results after incorporating social media features, especially for tracks with Medium and High popularity. While the results at this phase were more comparable to those of existing studies which attained accuracy, precision and recall scores above 80% [3,16], they are still slightly weaker, possibly due to the number of target classes predicted. There are three classes of popularity as opposed to the more commonly used binary classification seen in other studies. This is supported by the generally lower recall scores (ratio of correctly predicted observations to all the observations in the actual class) for the level 2 category (Medium popularity) relative to levels 1 and 3 (Low and High popularity), as illustrated in Table 10 . This indicates that it may be more challenging to identify the mediocre performing tracks as a distinct class while it is more clear-cut to differentiate hits from non-hits in a two-class classification.

Table 10. Comparisons of macro-recall for the four models

Models	Popularity Category	Length	Max	Sum	Mean	Debut
Random Forest	Low (Level 1)	96.0%	96.0%	96.0%	96.0%	96.0%
	Medium (Level 2)	63.6%	57.1%	58.3%	50.0%	30.0%
	High (Level 3)	71.4%	55.6%	84.6%	73.3%	73.3%
Ordinal Regression	Low (Level 1)	88.0%	80.0%	80.0%	76.0%	76.0%
	Medium (Level 2)	54.5%	57.1%	58.3%	60.0%	40.0%
	High (Level 3)	50.0%	44.4%	61.5%	60.0%	46.7%
Support Vector Machine	Low (Level 1)	96.0%	84.0%	84.0%	88.0%	84.0%
	Medium (Level 2)	36.4%	57.1%	50.0%	70.0%	50.0%
	High (Level 3)	57.1%	50.0%	69.2%	66.7%	46.7%
Neural Network	Low (Level 1)	64.0%	92.0%	96.0%	92.0%	92.0%
	Medium (Level 2)	18.2%	28.6%	8.3%	20.0%	10.0%
	High (Level 3)	0%	72.2%	76.9%	86.7%	73.3%

Notwithstanding, relatively good prediction results proved the effectiveness of these factors for hit songs prediction. On top of that, it implies that social media data may have enough discriminating power for separating hits and flops despite the shortcomings caused by small datasets. Indeed, several prominent studies that used limited sample sizes in the prediction of movie box performances managed to yield results sufficient to draw reasonable conclusions that justified the use of social media features [11,25]. These findings also bear implications on the usage of social media data, in particular YouTube features, for future research initiatives.

This study advocates the identification of potential popular songs post-release by collecting and tracking data from specific social media platforms and mediums. As opposed to the lack of existing literature that tapped onto social media listening behaviour, monitoring of such information is more ubiquitous in real-life applications. For instance, there is an abundance of music analytics tools such as Soundcharts and Chartmetric that help music professionals to aggregate real-time data, including social media audience data on multiple digital platforms, and radio airplay data for them to capitalize and take relevant actions.

5.3 Importance of audio and social media features

The top five variables were identified based on their variable importance scores to examine the influential features that impact the prediction in both stages. It is noted that for most of the models, each predictor variable got separate variable importance for each class. Support vector machine model has been used as the reference as it is the best performing model at the first stage.

Danceability, loudness and instrumentalness are among the significant determinants for the prediction of hit songs with audio features, although it is challenging to demonstrate their credibility given their subpar results. Nevertheless, these findings are partially in line with some current studies on HSS. For instance, danceability is a salient factor in predicting hit songs [1,13]. It is somewhat surprising that valence or in more layman terms, "happiness" of a song is not found to be influential for this dataset while found significant in other studies [1,13].

On the other hand, social media features, especially view counts, proved to be promising in predicting songs' popularity as per the variable importance of the random forest model. The results are fairly reasonable given that view counts in the first week of release for YouTube music videos signal the initial interest in the songs. Although comparisons cannot be made as no literature utilizes YouTube video metrics yet, similar studies have shown that listening behaviour enhanced prediction results [7].

In the movie industry, YouTube-based information of movie trailers served as an effective indicator of consumers' intention in watching the movie and thus, was positively correlated with box office gross revenue[11,25]. This also highlights the wide range of applications for data extracted from YouTube in various domains.

Other social media features such as votes per thousand views as well as dislike proportion were also among the top variables, albeit with much lower significance scores relative to view counts. A possible explanation for the lack of significance of these factors is that these music videos invited less voting or commenting from the audience, which renders these user engagement metrics less informative for prediction purposes. This interesting phenomenon was highlighted in the study by Liikkanen & Salovaara[19]. According to the study, while the mean views for videos belonging to the Music category were about 13 times larger compared to other genres including entertainment, pets, and animals as well as gaming, music videos were much less frequently commented and voted. The discovery led the authors to conclude that YouTube supports passive music listening with lesser user engagement activities and thus, functions more like a music streaming service.

6 Conclusions

To recap, this research adopts both audio features and social media variables, which are obtained from Spotify and YouTube music videos respectively for the prediction of songs' popularity. These platforms are apt given their high relevance and popularity in music streaming.

1. Random forest is the best performing model to predict music popularity for this dataset obtained the highest values for all four evaluation metrics. On average, the model generated accuracy of 79.6%, macro-precision of 74.5%, macro-recall of 73.2%, and macro F1-scores of 73.1%.
2. More importantly, this research has demonstrated that the inclusion of YouTube-based social media features is effective for the prediction of potential hit songs, together with audio features. This is justified by the improvement in the evaluation metrics, which ranged from 10% to 60%. In particular, macro F1-scores saw the largest improvement of 39% on average across all four models.

7 Limitations and Challenges

The findings of the study need to be interpreted in light of several limitations about the methodology used. First, there is a demographic gap between the users' base of Spotify and YouTube. The majority of Spotify users are from Europe and North America while YouTube users are more concentrated in other parts of the world. Nevertheless, this problem is not perceived to be of major significance given that the tracks are mostly in the English language, which is typically favoured and accepted by global audiences. Only a few tracks in the dataset were identified in non-English languages such as Spanish, Korean, and Latin.

Furthermore, another possible limitation was the time bias arising from the collection of tracks that have different release dates. Intuitively, songs released earlier may stay on the chart longer than songs released at later dates. In other words, these tracks may tend to fall in the High popularity category for dependent variables Length and Sum in particular, which are both affected by the period in which a song appeared in a chart. Nonetheless, a quick inspection of the tracks in the Medium and High popularity organized by their release dates revealed that this issue may not be critical as a subset of tracks released in August still managed to achieve High popularity across all target variables.

Apart from the limitations, the challenges deserve a highlight as current researchers in similar initiatives may face. The main obstacle lies in obtaining the data from multiple sources and merging them to derive a dataset, given that no such dataset which combines both audio and social media features is publicly available. Furthermore, social media variables are collected using YouTube Data API, which imposes a specified daily quota allocation for its users. The arduous data gathering process might have discouraged studies attempting to utilize social media variables, especially YouTube features for HSS.

8 Recommendations for Future Work

Given the relative lack of previous literature, this research is considered as an exploratory pilot study to investigate the potential application of YouTube-based features for hit song predictions. There is yet plenty possibly be done for future research. A larger dataset can potentially boost the overall performance scores. Furthermore, future directions can focus on exploring the use of more variety of YouTube-based features, especially those that describe the early popularity evolution pattern of the music videos and the social influence of the artist on the social media channel. Examples of such features are view count increase of the YouTube music video on the n^{th} day after the song is released, subscriber count of the artist on YouTube and others inspired by the study[26].

References

1) Ovide S. YouTube Isn't the Music Villain Anymore. 2021. Available from: https://www.nytimes.com/2021/06/08/technology/youtube-music-industry.html.
2) Aniftos R. Here Are YouTube's 10 Most-Watched Music Videos. . Available from: https://www.billboard.com/articles/news/9473980/youtube-most-watched-music-videos/.
3) Ahmad IS, Bakar AA, Yaakub MR. Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews. *Information Processing & Management*. 2020;57(5):102278–102278. Available from: https://doi.org/10.1016/j.ipm.2020.102278.
4) Watts D, Hasker S. Marketing in an unpredictable world. 2006. Available from: https://hbr.org/2006/09/marketing-in-an-unpredictable-world.
5) Interiano M, Kazemi K, Wang L, Yang J, Yu Z, Komarova NL. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*. 2018;5(5):171274–171274. Available from: https://royalsocietypublishing.org/doi/10.1098/rsos.171274.
6) Kim ST, Oh JH. Music intelligence: Granular data and prediction of top ten hit songs. *Decision Support Systems*. 2021;145:113535–113535. Available from: https://doi.org/10.1016/j.dss.2021.113535.
7) Middlebrook K, Sheik K. Song hit prediction: Predicting billboard hits using Spotify data. . Available from: https://arxiv.org/pdf/1908.08609.
8) Zangerlee, Vötter M, Huber R, Yang YH. Hit Song Prediction: Leveraging Low-and High-Level Audio Features. In: Proceedings of the 20th ISMIR Conference. 2019;p. 319–326. Available from: https://archives.ismir.net/ismir2019/paper/000037.pdf.
9) Martin-Gutierrez D, Penaloza GH, Belmonte-Hernandez A, Garcia FA. A Multimodal End-to-End Deep Learning Architecture for Music Popularity Prediction. *IEEE Access*. 2020;8:39361–39374. Available from: https://doi.org/10.1109/ACCESS.2020.2976033.
10) Singhi A, Brown DG. Can song lyrics predict hits. *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research 2015 P 457-471*. Available from: https://cs.uwaterloo.ca/~browndg/CMMR15data/CMMR2015paper.pdf.
11) Herremans D, Bergmans T. Hit song prediction based on early adopter data and audio features. . Available from: https://arxiv.org/pdf/2010.09489.
12) Zangerle E, Pichl M, Hupfauf B, Specht G. Can Microblogs Predict Music Charts? An Analysis of the Relationship Between# Nowplaying Tweets and Music Charts. *Proceedings of the 17th ISMIR Conference*. 2011. Available from: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/039_Paper.pdf.
13) Kim ST, Oh JH. Music intelligence: Granular data and prediction of top ten hit songs. *Decision Support Systems*. 2021;145:113535–113535. Available from: https://doi.org/10.1016/j.dss.2021.113535.
14) Middlebrook K, Sheik K. Song hit prediction: Predicting billboard hits using spotify data. 2019. Available from: https://arxiv.org/pdf/1908.08609.
15) Lee J, Lee JS. Music Popularity: Metrics, Characteristics, and Audio-Based Prediction. *IEEE Transactions on Multimedia*. 2018;20(11):3173–3182. Available from: https://doi.org/10.1109/TMM.2018.2820903.
16) Araujo CS, Cristo M, Giusti R. Predicting Music Popularity on Streaming Platforms. *Anais do Simpósio Brasileiro de Computação Musical (SBCM 2019)*. 2020;27:108–117.
17) Reference. Spotify for Developers. . Available from: https://developer.spotify.com/documentation/web-api/reference/.
18) Liikkanen LA. Three Metrics for Measuring User Engagement with Online Media and a YouTube Case Study. . Available from: https://arxiv.org/ftp/arxiv/papers/1312/1312.5547.
19) Liikkanen LA, Salovaara A. Music on YouTube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior*. 2015;50:108–124. Available from: https://doi.org/10.1016/j.chb.2015.01.067.
20) Breiman L. Random forests. *Machine Learning*. 2001;45:5–32. Available from: https://doi.org/10.1023/A:1010933404324.
21) Biau G, Scornet E. A random forest guided tour. *TEST*. 2016;25(2):197–227. Available from: https://doi.org/10.1007/s11749-016-0481-7.
22) Opitz J, Burst S. Macro F1 and macro F1. . Available from: https://arxiv.org/pdf/1911.03347.pdf.
23) Savage M. Pop music is getting faster (and happier). . Available from: https://www.bbc.com/news/entertainment-arts-53167325.
24) Raza AH, Nanath K. Predicting a Hit Song with Machine Learning: Is there an apriori secret formula? *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*. 2020;p. 111–116. Available from: https://doi.org/10.1109/DATABIA50434.2020.9190613.
25) Oh C, Roumani Y, Nwankpa JK, Hu HFF. Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Information & Management*. 2017;54(1):25–37. Available from: https://doi.org/10.1016/j.im.2016.03.004.
26) Ouyang S, Li C, Li X. A Peek Into the Future: Predicting the Popularity of Online Videos. *IEEE Access*. 2016;4:3026–3033. Available from: https://doi.org/10.1109/ACCESS.2016.2580911.