CSE343 : ML Project Monsoon 2021

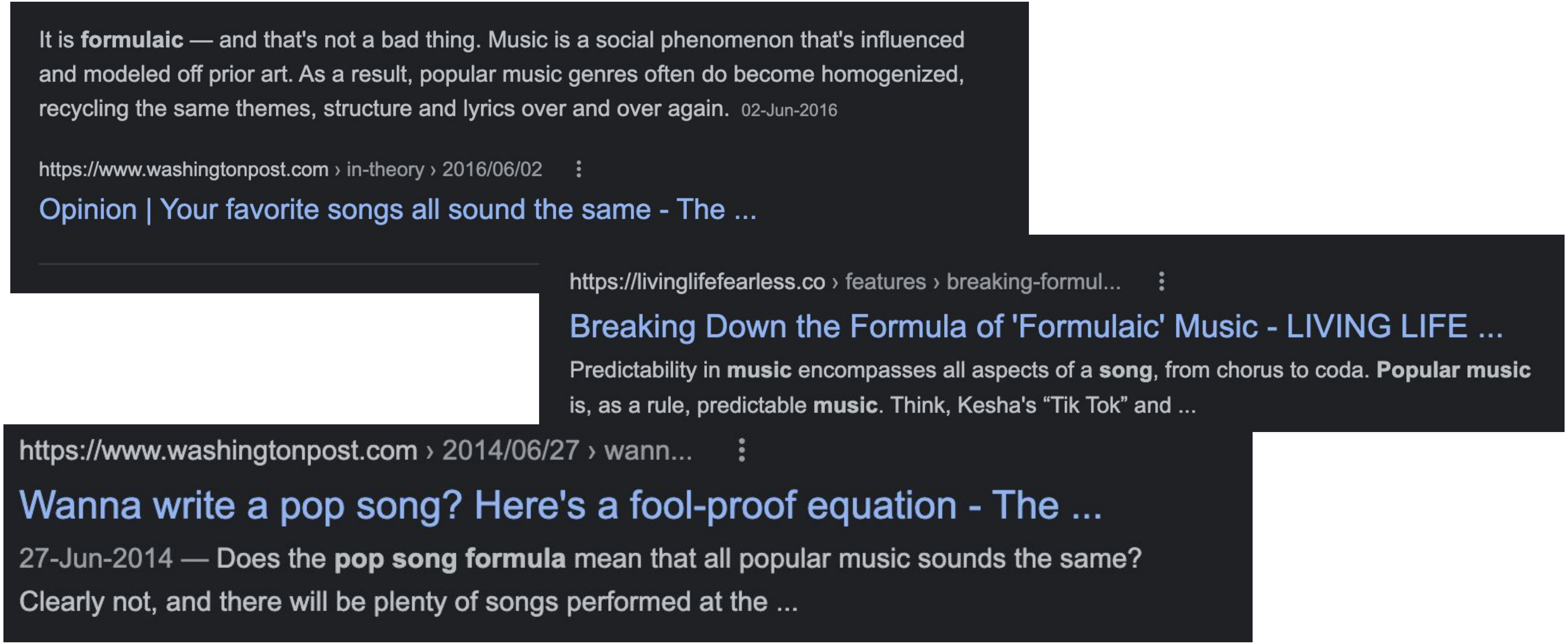
Ansh Arora (2019022)
Jishnu Raj Parashar (2019048)
Nandika Jain (2019064)
Tushar Mohan (2019393)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

OK Computer: Employing Machine Learning to analyse the impact of different parameters in foretelling the next earworm

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation



It is **formulaic** — and that's not a bad thing. Music is a social phenomenon that's influenced and modeled off prior art. As a result, popular music genres often do become homogenized, recycling the same themes, structure and lyrics over and over again. 02-Jun-2016

https://www.washingtonpost.com › in-theory › 2016/06/02

Opinion | Your favorite songs all sound the same - The ...

https://livinglifefearless.co › features › breaking-formul...

Breaking Down the Formula of 'Formulaic' Music - LIVING LIFE ...

Predictability in **music** encompasses all aspects of a **song**, from chorus to coda. **Popular music** is, as a rule, predictable **music**. Think, Kesha's "Tik Tok" and ...

https://www.washingtonpost.com › 2014/06/27 › wann...

Wanna write a pop song? Here's a fool-proof equation - The ...

27-Jun-2014 — Does the **pop song formula** mean that all popular music sounds the same? Clearly not, and there will be plenty of songs performed at the ...

Motivation

- In the age of TikTok and Reels, it is claimed that the songs regularly reaching the top of the charts are getting repetitive, almost following a pattern that guarantees its success.
- Some artists (Another one, DJ KHALED!) have even gone on to make rather braggadocious claims that they know well before the release of their song whether it's going to be a hit or not.
- We wish to answer these burning questions - What is it, after all, that seems to make a song more popular than others, and can we accurately predict the popularity of a track.

Literature Review

Hit Song Science Once Again a Science by Ni et al.

- This paper talks about using the shifting perceptron algorithm to classify top 5 songs from the rest from the weekly charts over the last 50 years.
- Features like tempo, time signature, song duration and loudness, coefficient of variance of loudness and harmonic simplicity were used.
- The EchoNest API has been used to extract the features for the 5947 unique songs that were collected from the Official Charts Company.
- Successfully investigates the change in musical taste over the years which separates it from its previous studies. Uses classifiers only.

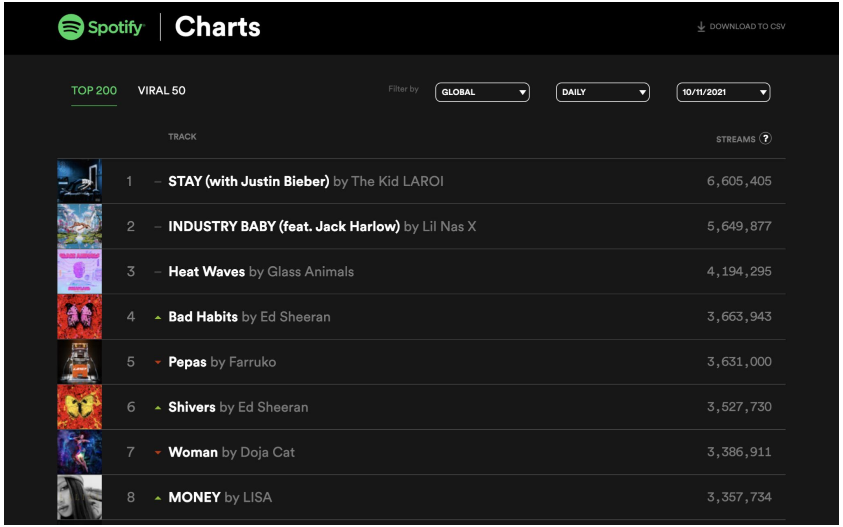
Literature Review

- ***Predicting Music Popularity on Streaming Platforms*** by Araujo et al.

- This paper talks about the methodology to predict if a song will appear on Spotify’s Top 50 Global ranking after a certain amount of time.
- Features such as entry’s rank, duration, explicit flag and daily popularity score were used for the same.
- SVM, Random Forests and Gaussian Na ̈ive-Bayes models were used to classify whether a song is successful or not based on its appearance on the Spotify Top 50 Global charts.

Dataset

We were after the latest possible data available for our tests. Thus we set out to create our own dataset and made use of Spotify Charts data from 2017-21, which we procured from the Spotify Charts site.

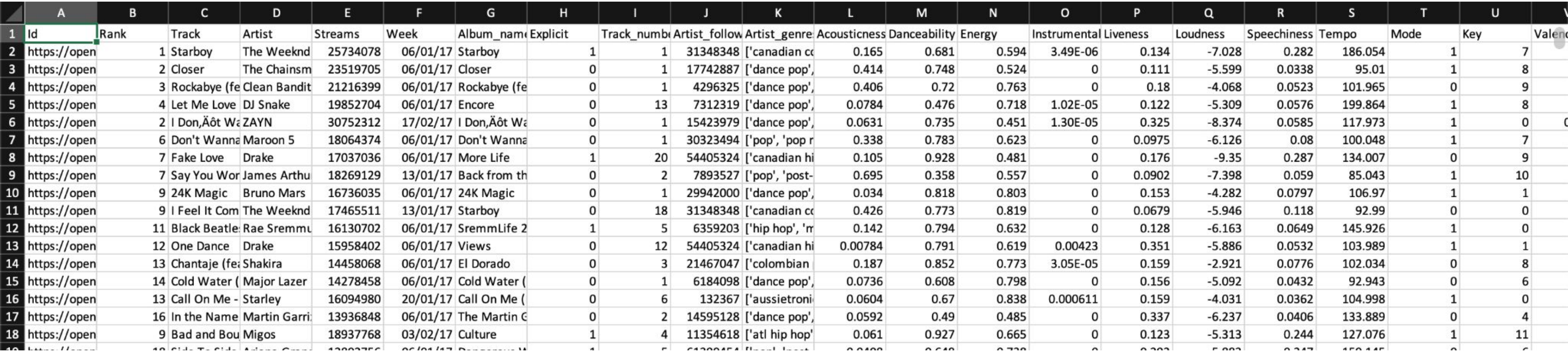


Filter by GLOBAL DAILY 10/11/2021

		TRACK	STREAMS
1	–	**STAY (with Justin Bieber)** by The Kid LAROI	6,605,405
2	–	**INDUSTRY BABY (feat. Jack Harlow)** by Lil Nas X	5,649,877
3	–	**Heat Waves** by Glass Animals	4,194,295
4	▲	**Bad Habits** by Ed Sheeran	3,663,943
5	▼	**Pepas** by Farruko	3,631,000
6	▲	**Shivers** by Ed Sheeran	3,527,730
7	▼	**Woman** by Doja Cat	3,386,911
8	▲	**MONEY** by LISA	3,357,734

Dataset

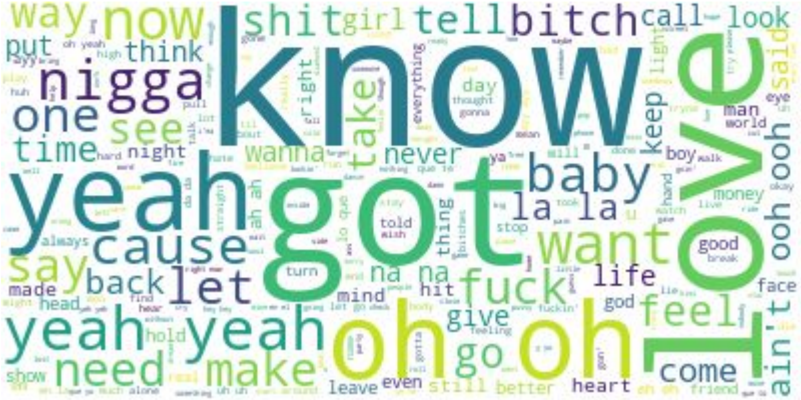
The Spotify ID was then used with the Spotify Web API to garner the set of acoustic features that it provides. As a result, we were able to generate a dataset of 4247 songs, each with a total of 23 features -

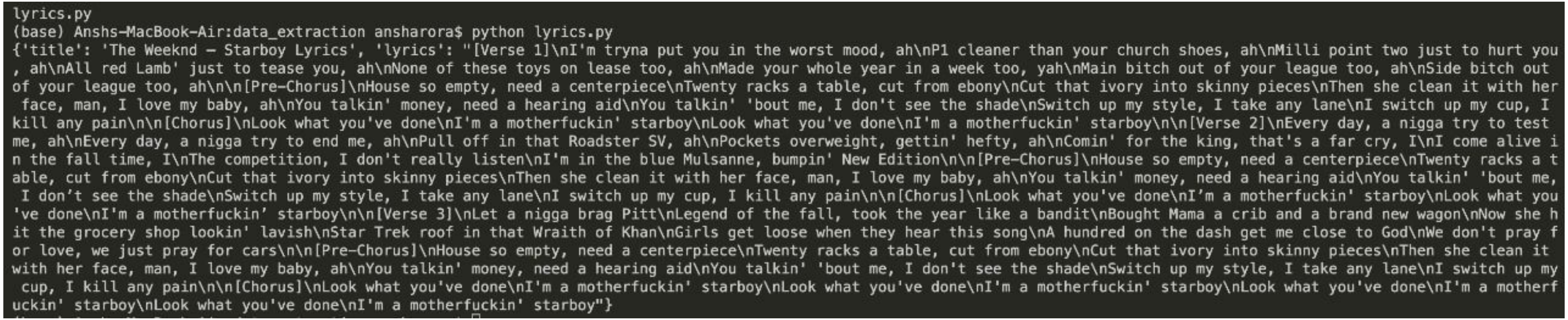


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Id	Rank	Track	Artist	Streams	Week	Album_name	Explicit	Track_numbe	Artist_follow	Artist_genre	Acousticness	Danceability	Energy	Instrumental	Liveness	Loudness	Speechiness	Tempo	Mode	Key	Valenc
2	https://open	1	Starboy	The Weeknd	25734078	06/01/17	Starboy	1	1	31348348	['canadian cc	0.165	0.681	0.594	3.49E-06	0.134	-7.028	0.282	186.054	1	7	
3	https://open	2	Closer	The Chainsm	23519705	06/01/17	Closer	0	1	17742887	['dance pop',	0.414	0.748	0.524	0	0.111	-5.599	0.0338	95.01	1	8	
4	https://open	3	Rockabye (fe	Clean Bandit	21216399	06/01/17	Rockabye (fe	0	1	4296325	['dance pop',	0.406	0.72	0.763	0	0.18	-4.068	0.0523	101.965	0	9	
5	https://open	4	Let Me Love	DJ Snake	19852704	06/01/17	Encore	0	13	7312319	['dance pop',	0.0784	0.476	0.718	1.02E-05	0.122	-5.309	0.0576	199.864	1	8	
6	https://open	2	I Don,Äôt Wa	ZAYN	30752312	17/02/17	I Don,Äôt Wa	0	1	15423979	['dance pop',	0.0631	0.735	0.451	1.30E-05	0.325	-8.374	0.0585	117.973	1	0	0
7	https://open	6	Don't Wanna	Maroon 5	18064374	06/01/17	Don't Wanna	0	1	30323494	['pop', 'pop r	0.338	0.783	0.623	0	0.0975	-6.126	0.08	100.048	1	7	
8	https://open	7	Fake Love	Drake	17037036	06/01/17	More Life	1	20	54405324	['canadian hi	0.105	0.928	0.481	0	0.176	-9.35	0.287	134.007	0	9	
9	https://open	7	Say You Wor	James Arthu	18269129	13/01/17	Back from th	0	2	7893527	['pop', 'post-	0.695	0.358	0.557	0	0.0902	-7.398	0.059	85.043	1	10	
10	https://open	9	24K Magic	Bruno Mars	16736035	06/01/17	24K Magic	0	1	29942000	['dance pop',	0.034	0.818	0.803	0	0.153	-4.282	0.0797	106.97	1	1	
11	https://open	9	I Feel It Com	The Weeknd	17465511	13/01/17	Starboy	0	18	31348348	['canadian cc	0.426	0.773	0.819	0	0.0679	-5.946	0.118	92.99	0	0	
12	https://open	11	Black Beatle:	Rae Sremmu	16130702	06/01/17	SremmLife 2	1	5	6359203	['hip hop', 'm	0.142	0.794	0.632	0	0.128	-6.163	0.0649	145.926	1	0	
13	https://open	12	One Dance	Drake	15958402	06/01/17	Views	0	12	54405324	['canadian hi	0.00784	0.791	0.619	0.00423	0.351	-5.886	0.0532	103.989	1	1	
14	https://open	13	Chantaje (fe:	Shakira	14458068	06/01/17	El Dorado	0	3	21467047	['colombian	0.187	0.852	0.773	3.05E-05	0.159	-2.921	0.0776	102.034	0	8	
15	https://open	14	Cold Water (	Major Lazer	14278458	06/01/17	Cold Water (	0	1	6184098	['dance pop',	0.0736	0.608	0.798	0	0.156	-5.092	0.0432	92.943	0	6	
16	https://open	13	Call On Me -	Starley	16094980	20/01/17	Call On Me (	0	6	132367	['aussietroni	0.0604	0.67	0.838	0.000611	0.159	-4.031	0.0362	104.998	1	0	
17	https://open	16	In the Name	Martin Garri:	13936848	06/01/17	The Martin G	0	2	14595128	['dance pop',	0.0592	0.49	0.485	0	0.337	-6.237	0.0406	133.889	0	4	
18	https://open	9	Bad and Bou	Migos	18937768	03/02/17	Culture	1	4	11354618	['atl hip hop'	0.061	0.927	0.665	0	0.123	-5.313	0.244	127.076	1	11	

Dataset

We extracted the lyrics using Google's Engine API. And then performed Topic Modelling using LDA to extract features from the Lyrics Dataset.



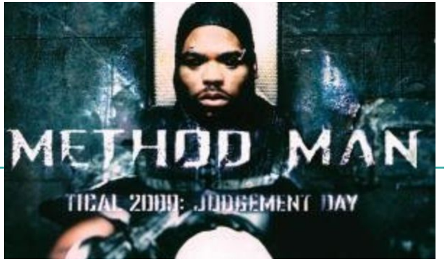


```
lyrics.py
(base) Anshs-MacBook-Air:data_extraction ansharora$ python lyrics.py
{'title': 'The Weeknd – Starboy Lyrics', 'lyrics': "[Verse 1]\nI'm tryna put you in the worst mood, ah\nP1 cleaner than your church shoes, ah\nMilli point two just to hurt you
, ah\nAll red Lamb' just to tease you, ah\nNone of these toys on lease too, ah\nMade your whole year in a week too, yah\nMain bitch out of your league too, ah\nSide bitch out
of your league too, ah\n\n[Pre-Chorus]\nHouse so empty, need a centerpiece\nTwenty racks a table, cut from ebony\nCut that ivory into skinny pieces\nThen she clean it with her
 face, man, I love my baby, ah\nYou talkin' money, need a hearing aid\nYou talkin' 'bout me, I don't see the shade\nSwitch up my style, I take any lane\nI switch up my cup, I
kill any pain\n\n[Chorus]\nLook what you've done\nI'm a motherfuckin' starboy\nLook what you've done\nI'm a motherfuckin' starboy\n\n[Verse 2]\nEvery day, a nigga try to test
me, ah\nEvery day, a nigga try to end me, ah\nPull off in that Roadster SV, ah\nPockets overweight, gettin' hefty, ah\nComin' for the king, that's a far cry, I\nI come alive i
n the fall time, I\nThe competition, I don't really listen\nI'm in the blue Mulsanne, bumpin' New Edition\n\n[Pre-Chorus]\nHouse so empty, need a centerpiece\nTwenty racks a t
able, cut from ebony\nCut that ivory into skinny pieces\nThen she clean it with her face, man, I love my baby, ah\nYou talkin' money, need a hearing aid\nYou talkin' 'bout me,
 I don't see the shade\nSwitch up my style, I take any lane\nI switch up my cup, I kill any pain\n\n[Chorus]\nLook what you've done\nI'm a motherfuckin' starboy\nLook what you
've done\nI'm a motherfuckin' starboy\n\n[Verse 3]\nLet a nigga brag Pitt\nLegend of the fall, took the year like a bandit\nBought Mama a crib and a brand new wagon\nNow she h
it the grocery shop lookin' lavish\nStar Trek roof in that Wraith of Khan\nGirls get loose when they hear this song\nA hundred on the dash get me close to God\nWe don't pray f
or love, we just pray for cars\n\n[Pre-Chorus]\nHouse so empty, need a centerpiece\nTwenty racks a table, cut from ebony\nCut that ivory into skinny pieces\nThen she clean it
with her face, man, I love my baby, ah\nYou talkin' money, need a hearing aid\nYou talkin' 'bout me, I don't see the shade\nSwitch up my style, I take any lane\nI switch up my
 cup, I kill any pain\n\n[Chorus]\nLook what you've done\nI'm a motherfuckin' starboy\nLook what you've done\nI'm a motherfuckin' starboy\nLook what you've done\nI'm a motherf
uckin' starboy\nLook what you've done\nI'm a motherfuckin' starboy"}
```

Methodology

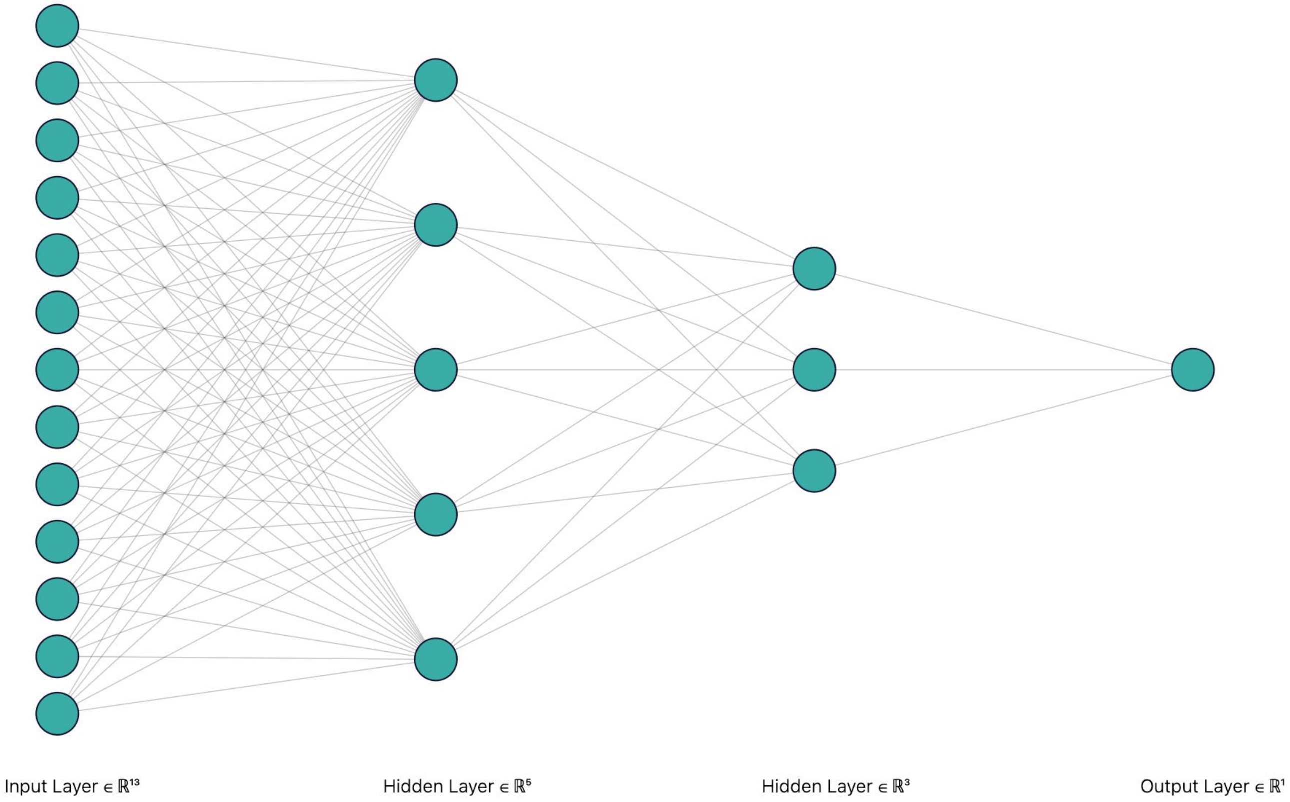
- We created labels for our use cases - For Classification Problems, we used Hit (1) and Not-Hit (0) and for Regression Problems, we used a gradient scale from 0 to 1, 1 for songs that have charted at the top, 0.1 for songs that peaked at #200 and 0 for songs that have not charted yet.
- We carried out Classification Algorithms - **Logistic Regression** and **Support Vector Machines** and have made use of Regression Algorithms - **Logistic Regression, Regularized Regression** and **Support Vector Regression**. We utilized **Cross Validation techniques** such as **K-Fold** and Holdout, and made use of GridSearchCV as well in cases to get the best hyperparameters.

Methodology



- We further experimented with binary and multi-class classification and settled on binary classification as it came out to be more accurate for our experiment. We performed cross validation methods on **Random Forests** and Boosting techniques **Adaboost** and **XGBoost**.
- To bolster our classification performance further, we made use of Keras' **Artificial Neural Networks**. We tested our data on NNs with 1 and 2 hidden layers over various activation functions and epochs.

Methodology


Input Layer $\in \mathbb{R}^{13}$
Hidden Layer $\in \mathbb{R}^{5}$
Hidden Layer $\in \mathbb{R}^{3}$
Output Layer $\in \mathbb{R}^{1}$

Methodology

Clustering

K-Means clustering was tested with different initial clusters and their respective silhouette scores.

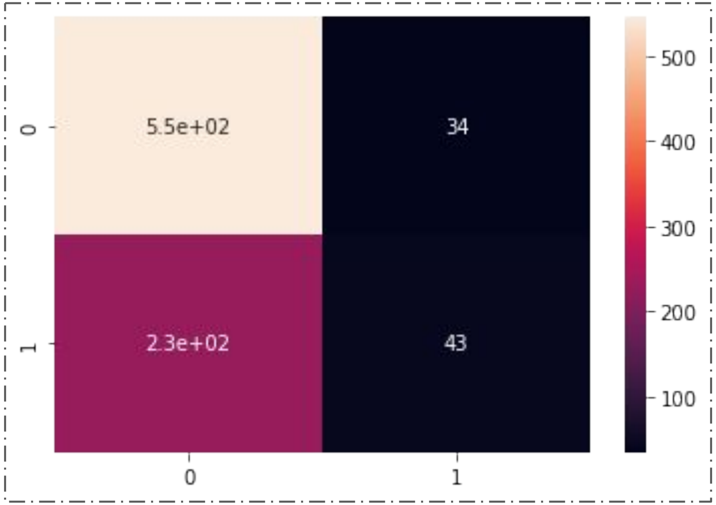
Sentiment Analysis

VADER and **TextBlob** were 2 techniques utilized for extracting more features from the dataset of lyrics. Former has is it roots to lexicon-based analysis while the latter one relies on pattern-matching analysis.

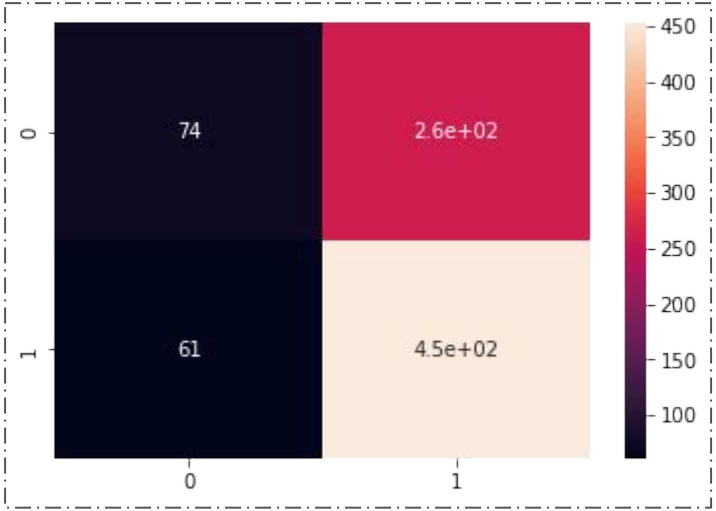
Results and Analysis

From the various models that we applied on our dataset, we found almost negligible correlation between any 2 features. Also, our logistic regression performs fairly well with accuracy peaking at **71%** and precision at **84%**. On apply LDA, we could extract **18** features with the best coherence score out of the **27** different range of features we tested on. Linear regression gave an average RMSE of **0.039**. Regularized Lasso Regression returned slightly better values here but not as good as Ridge Regression. Also, on doing the t-test and f-test on Linear regression model, we confirmed that our assumption of residual errors being normally distributed with mean 0 stands strong as the p-value calculated is above 5%.

Results and Analysis



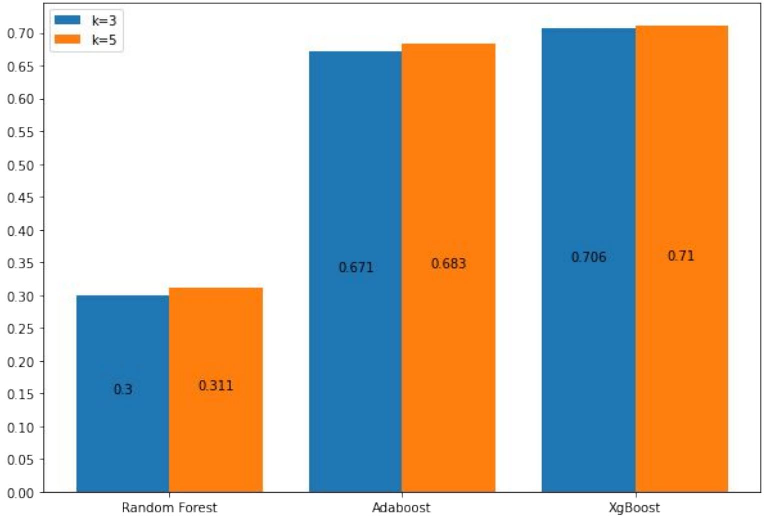
Keeping a threshold of 50, and kernel as linear.



Keeping a threshold of 100, and kernel as poly.

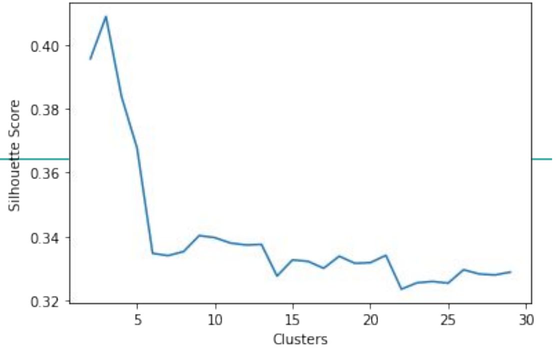
Results and Analysis

Bagging techniques like Random Forest allowed us to obtain a mean accuracy of **31.1%**. Boosting techniques like XGBoost and Adaboost performed better achieving a maximum mean accuracy of **71% & 68.3%** on a cross-validation test.



ANNs returned a decent CV score of **73.5%** with ReLU activation function on a single hidden layer and **73.8%** on two hidden layers. Other activation functions and hyperparameters performed in a similar ballpark, returning accuracies around **71%.** This was in clear contrast to Keras' baseline Logistic Regression which gave a max accuracy of **68.8%**.

Results and Analysis

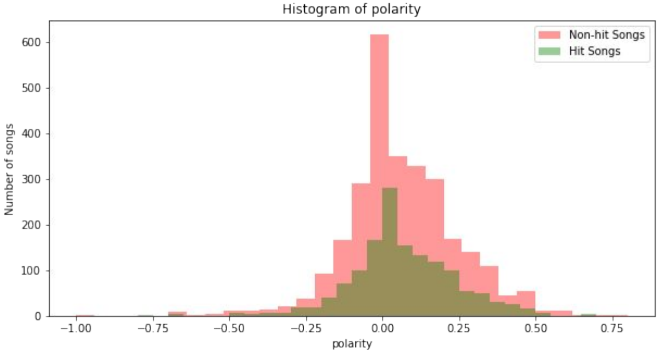


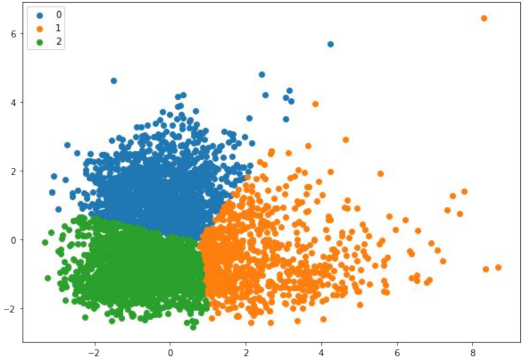
Clustering:

3 clusters gave the highest silhouette score so we chose that and visualised on a 2D plane using **PCA**.

Sentiment Analysis:

Most songs in the top charts tend to be of neutral nature as shown in the histogram on the left.





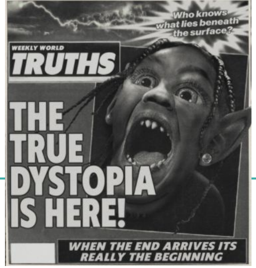
Future Work

Following the work already done, the accuracy from basic machine learning techniques is not good enough, so the subsequent work will be focused on **improving** our **multi-class classification** accuracy via advanced machine learning and deep learning classifiers.

We also wish to inculcate our models into a UI-based environment with **human-in-the-loop** approach to keep improving the models in the backend. This way we will be able to achieve fairly better accuracy as the dataset will no longer be restricted and the model will keep on learning.

The idea of topic modelling can be extended towards a linear combination approach to classify a song as a hit or a non-hit.

Conclusion


Who knows what lies beneath the surface?
WEEKLY WORLD
TRUTHS
THE TRUE DYSTOPIA IS HERE!
WHEN THE END ARRIVES ITS REALLY THE BEGINNING

We are pleased to inform that our experiments with various models, other than helping us better our understanding of applied Machine Learning models, have also given us good results. We were able to create a **dataset from scratch**, using data that hasn't been used in any study before. To perform supervised learning off of it, we created multiple classes of labels, including a numerical **HitScore** and simple **Binary (Hit - Non Hit)** and **Multi-Class** classification (Star Rating). While different models performed variably, we were able to get much better results than baseline in almost all cases, showing that there in fact is some truth to the fact that the songs that have been getting popular in the last 5 years have some amount of **common features to them acoustically**. Lyrically also our analysis of the polarity of lyrics showed that instead of getting more polarised songs on top of charts, **neutral songs** generally tend to be **hits**, further **reinforcing** our claims on **repetitiveness** being present in the world of **music**.

Contributions

Ansh	Lyrics Curation, Exploratory Data Analysis, Support Vector Regression, Artificial Neural Network, Clustering, Report + Powerpoint
Jishnu	Data Curation, Linear Regression, Logistic Regression, Support Vector Classification, Random Forest, Report + Powerpoint
Nandika	Data Curation, Linear Regression, Support Vector Classification, Logistic Regression, Random Forest, Report + Powerpoint
Tushar	Lyrics Curation, Exploratory Data Analysis, Natural Language Processing, Clustering, Report + Powerpoint

LINKIN PARK
IN THE END