Music Emotion Recognition with Audio and Lyrics Features

C. V. Nanayakkara[1] and H. A. Caldera[2]
University of Colombo School of Computing (UCSC)
35, Reid Avenue, Colombo 7, Sri Lanka
[1]charini.nanayakkara@gmail.com
[2]hac@ucsc.cmb.ac.lk

ABSTRACT

Music Emotion Recognition (MER) is a field of science dedicated to recognizing emotions associated with music pieces. With the new interest in music therapy and music recommendation systems, MER has caught immense interest of scientists. This study is an attempt at discerning how well music related emotions can be predicted with music features; audio and lyrics. Emotion classes associated with songs were initially identified with clustering. Independent classification experiments were executed utilizing lyrics and audio features, to assess the comparative best model for predicting music emotions. The classification algorithms attempted in this research are Naïve Bayes, Random Forest, SVM and C4.5 decision. Random Forest with oversampling on the audio feature set produced comparative best results.

KEYWORDS

Music Emotion Prediction, Lyrics and Audio Features, Machine Learning, Hierarchical Clustering, Classification, Music Specific Emotion Model

1 INTRODUCTION

The capability people possess to intuitively interpret the notion conveyed by a music piece, even in the absence of words, is possibly accountable to the omnipresence of music in our lives since birth. In fact, research attest to the fact that sensitivity to music is expressed even during prenatal stage, whereas the capability of perceiving emotion in music develops since infancy. Such extensive evidence concerning the impact which music depicts on emotional aspects have motivated researchers to conduct number of studies on the matter during recent years.

1.1 Music and Emotion: A Background Review

Music, according to the definition provided by WordNet, is “an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner” [1]. Three primary emotion categories have been identified to be associated with the context of music [2]. *Expressed emotion* relates to the emotions a performer or composer wishes to communicate to listeners through a song whereas *Felt emotion* reflects emotion actually felt by the listener when listening to music. However it’s *Perceived emotion* which is highlighted as main focus in Music Emotion Recognition (MER) context. This relates to the emotions a listener perceives or assumes as being expressed by a music composition.

1.2 Music Features

Four primary aspects; structural features, listener features, performance features and contextual features; are said to be decisive of what emotions an individual gets when listening to music [3]. Structural features relate to the sounds with which a song comprises of. Melody, tempo and rhythm are few of these structural features associated with music. It is inclusive of the lyrics of the song, which refers to the words a music piece comprises of. Listener features reflect on traits of each individual listener, such as age, reason for listening to a song, psychological state, etc. Mannerism in which the music piece is performed, skill and appearance of performer, etc. deal with performance features whereas contextual features relate to the location where music is performed (e.g. wedding, musical show, funeral). Among these, structural features and listener features (listener’s opinion of the emotions associated with the song) have been utilized in this research. The focus has been to conduct experiments to analyze the possibility of predicting musically induced emotions, with lyrics and audio features.

1.3 Research Problem

The primary research problem addressed in this paper is *evaluating capability of automatically predicting the emotions associated with a music piece, with a fair level of accuracy*. Emotions perceived by an individual when listening to a certain song, is referred to as 'emotions associated with a music piece' in this study. This research problem would be addressed by modelling several data driven approaches for prediction of musically induced emotion and evaluating them to ascertain their level of acceptability. Thus, it's required to identify the classes of emotion which are elicited in people by music, since music may be expressive of merely some emotions but not all. Furthermore this is necessitated by the distinction of musically induced emotion from the general class of emotions [4]. Subsequently, a model which has the capability of automatically identifying the emotions expressed by a song must be identified. This model may utilize either lyrics or audio features, based on what appears to be most promising in the field of MER, according to our research.

2 METHODOLOGY

The methodology devises a music specific emotion model and evaluates the capability of different classification algorithms to predict the emotions expressed by a song. The research was based on the Million Song Dataset (MSD). MSD comprises of a million songs whereas tags, lyrics and audio features associated with these songs are provided by musiXmatch, last.fm and Vienna University of Technology respectively.

Figure 1 depicts the architectural view which comprises of four primary components. Phase 1 of the research is dedicated to evaluating and benchmarking several music datasets, to evaluate their aptitude for achieving the research objectives. Each dataset is benchmarked with relation to relevance, quantity and quality subsequent to which the most suitable dataset to utilize for research is determined. Preprocessing and feature retrieval tasks are executed in phase 2 whereas phase 3 deals with forming a music specific emotion model using tags. Identification of emotion related tags has been initially executed following which dimensionality of emotion space has been reduced via combining synonyms and executing hierarchical clustering. Subsequently, classification algorithms; Naïve Bayes [5], Random Forest [6], Support Vector Machine (SVM) [7] and C4.5 [8]; have been attempted in phase 4. This last phase is executed separately using audio and lyric features.

3 DATA ACQUISITION

Six popular music datasets; RWC database, GTZAN genre collection, Uspop2002, MagnaTagATune, Musicbrainz and MSD; were benchmarked based on relevance, quantity and quality. While the magnitude of the former 3 datasets were inadequate considering the existence of larger music collections, audio features were required to be obtained separately for MagnaTagATune and Musicbrainz datasets. Considering the magnitude of dataset and availability of features necessary to conduct research, MSD was opted for.

The lyrics of songs in the MSD have been obtained from musiXmatch which is presently the largest lyrics catalogue in the world. Lyrics have been represented as bag-of-words in this dataset.

MSD provides with song level tags, which were utilized in the research for determining emotions associated with each song. Tags are 'terms' which song listeners have associated with the music pieces in the MSD, via the API provided by Last.fm . Due to this being a site used by millions of people around the world to satisfy their musical requirements, tags appearing in this dataset reflect the opinions of a global community, thus making it a suitable resource based on which to determine emotions perceived by music.

Furthermore, the Vienna University of Technology has provided with a multitude of audio features for 994 960 tracks in the MSD.

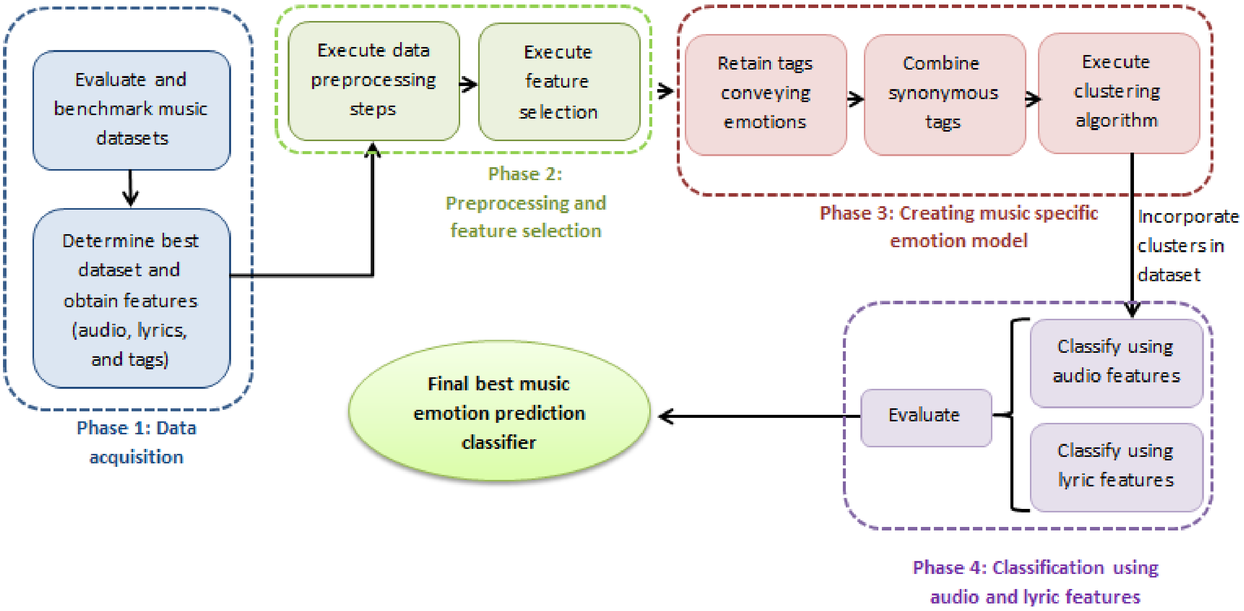

Evaluate and benchmark music datasets
Determine best dataset and obtain features (audio, lyrics, and tags)
Phase 1: Data acquisition
Execute data preprocessing steps
Execute feature selection
Phase 2: Preprocessing and feature selection
Retain tags conveying emotions
Combine synonymous tags
Execute clustering algorithm
Phase 3: Creating music specific emotion model
Incorporate clusters in dataset
Classify using audio features
Classify using lyric features
Evaluate
Final best music emotion prediction classifier
Phase 4: Classification using audio and lyric features

Figure 1: Overall architecture of proposed methodology

4 FEATURE ENGINEERING

Subsequent to selecting MSD via comparative analysis of several music datasets, preprocessing and feature selection were executed to refine dataset to suit classification task.

4.1 Data Preprocessing

Initially the data files were organized as ARFF and CSV files since utilized tools (R, Weka, RapidMiner, etc.) supported these formats.

The lyric words associated with songs were then weighted by Term Frequency – Inverse Document Frequency (TF-IDF) value. TF-IDF weighting helps determine the importance of a term in a document, with relation to a collection of documents considered [9]. This extensively applied score in document classification context could be utilized in our research by considering songs as ‘documents’ and words from song lyrics as ‘terms’. Rather than merely depicting the presence or absence of a word from song lyrics, using TF-IDF value allows reflecting how important a word is for a song. Calculation of TF-IDF score is as follows.

Term Frequency (TF) calculates the frequency with which a term occurs in a document [9] (1).

$$TF = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{count\ of\ all\ words\ occuring\ in\ document\ d} \quad (1)$$

Inverse Document Frequency (IDF) of a term is defined as follows. It assigns a high value to those terms which rarely occur, whereas words commonly appearing in a number of documents get a low value (2).

$$IDF = \frac{total\ number\ of\ documents\ in\ the\ corpus}{total\ number\ of\ documents\ with\ term\ t} \quad (2)$$

Final TF-IDF value of term t in document d is calculated by multiplying these two values; (1) and (2). It reflects importance of t with respect to d.

Using IDF together with TF helps give more importance to words that often occur in a given song, but less in the collection of music. Furthermore, it helps minimize the importance given to frequently occurring terms such as ‘the’ and ‘a’.

Using this measure, the importance of each lyric word with relation to a specific track was calculated. If TF-IDF of a word was less than the average TF-IDF value of all words for corresponding song, we considered the word to be of no significance for that song. Commonly occurring words were thus removed from the lyrics dataset. This reduced the lyric feature dimensionality from 5000 terms to 1536 English words. A CSV file corresponding to this lyric ARFF file was created, Of the one million songs provided in Million Song Dataset, audio, lyric and tag features were not provided for all. This has resulted in occurrence of missing values in the dataset. It was resolved via elimination of a tuple if a feature was missing since the resultant dataset was of considerable magnitude (167,023 songs).

Tags which conveyed emotion alone were identified and retained using WordNet tool and GEMS [4]. This preprocessing stage was important for creation of music specific emotion model. Outliers of the dataset were identified using the Inter Quartile Range (IQR).

4.2 Feature Selection

Feature selection is the mechanism of selecting a subset from feature space, which is most relevant to class attribute values. Feature redundancy analysis, otherwise known as correlation based feature selection is a renowned filter technique used for selecting subset of features. Linear correlation coefficient ρ, which is the primary measure adopted in this technique, is defined in (3), where $x_{i,}$ y_i are two variables and $\bar{x}$, $\bar{y}$ are their respective means. ρ is between -1 and 1 where equality to -1 or 1 signifies complete correlation. Features depicting high correlation are considered to be redundant, thus reflecting that retaining only one of such a feature pair is adequate for classification. Subsequent to applying this algorithm using (+/-) 0.75 as threshold, ten audio features were retained.

$$\rho = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \qquad (3)$$

Of the Audio features depicted in Table 1, mean and standard deviation values of Spectral centroid, Spectral flux, Compactness, Spectral variability and Fraction of low energy windows were retained subsequent to applying correlation. Subsequent to applying correlation based feature selection to the lyrics feature space, it was noted that no significant correlation was depicted among words in songs. Thus we were compelled to retain all 1536 lyric words for classification.

Table 1: Audio features

Tool		Feature	Description
Tools used by Vienna University for retrieving audio features	JMIR	Spectral centroid	The center of mass of the power spectrum.
		Spectral roll-off point	The fraction of bins in the power spectrum at which 85% of the power is at lower frequencies. This is a measure the right-skewedness of the power spectrum.
		Spectral flux	A measure of the amount of spectral change in a signal. Found by calculating the change in the magnitude spectrum from frame to frame.
		Compactness	A measure of the noisiness of a recording. Found by comparing the components of a window's magnitude spectrum with the magnitude spectrum of its neighboring windows.
		Spectral variability	The standard deviation of the magnitude spectrum.
		Root mean square	A measure of the power of a signal over a window.
		Zero crossings	The number of times the waveform changed sign in a window. An indication of frequency as well as noisiness.
	Marsyas	Fraction of low energy windows	The fraction of the last 100 windows that has an RMS less than the mean RMS of the last100 windows.
		Timbre features	Measurement on the Fast Fourier Transformation (FFT) of sounds generated by octave notes.

5 MUSIC EMOTION MODEL

In agglomerative hierarchical clustering approach, each data point (tags in this instance) would be in separate clusters at the beginning, which would be successively merged to form larger clusters. The merging could be halted either when all data points are in one cluster or when a certain stopping criterion is met [10]. Merging of two clusters in agglomerative hierarchical clustering is based on the linkage type opted for. It specifies the criterion of merging two or more clusters (i.e. whether to consider the furthest data points of two clusters, the closest, etc.). Furthermore, a distance measure is associated with linkage, which specifies the manner in which to measure distance between two (or more) clusters.

When applying hierarchical clustering in this research, the distance measure used was jaccard distance x [11] which is defined in (4). Five linkage methods single, complete, group average, Ward1 and Ward 2 [10] [12] were applied. Thus the final clusters were formed via combining tags which appear together at the lowest level of hierarchy in 80% of these algorithms. This is a significant deviation from the common method of opting for one linkage method over the rest. On the contrary, tags, which were assigned to the same cluster in 4 out of 5 (80%) clustering approaches attempted, were grouped together (e.g. if tag A and tag B were grouped together in 1st iteration of clustering using complete, group average, Ward1 and Ward2 methods, except in single linkage method, those two tags were still qualified for merging due to 80% or more constraint).

$$x = 1 - \frac{|songs\ with\ tag\ A \cap songs\ with\ tag\ B|}{|songs\ with\ tag\ A \cup songs\ with\ tag\ B|} \qquad (4)$$

The emotion clusters formed when this terminology was adhered to, are as depicted in Table 2. Emotion clusters have been numbered from C1 to C25. Column ‘Count’ specifies the number of songs which belong to each cluster whereas column ‘Emotion Tags’ conveys which tags each cluster comprises of. These emotion clusters were incorporated in the dataset prior to proceeding with supervised learning (i.e. If all the tags in a specific cluster were associated with a particular song, that cluster was incorporated with relevant song). Since more than 80% of the songs are distributed among seven majority emotion classes; C1, C2, C3, C7, C17, C22 and C25; merely those classes were retained for conducting classification experiments.

Table 2: 25 emotion clusters

Class	**Emotion tags**	**Count**
C1	Joyful, Danceable	1202
C2	Witty	1734
C3	Calming, Melancholic, Romantic	1639
C4	Uplifting, Feel good	199
C5	Aggressive, Angry	78
C6	Moving, Inspiring	20
C7	Sensual	3821
C8	Haunting, Dark, Depressing	16
C9	Angst	211
C10	Sick	177
C11	Makes me smile	340
C12	Heartbreaking, Bittersweet	23
C13	Crazy	251
C14	Exciting	175
C15	Creepy	371
C16	Ethereal, Dreamy, Soothing	7
C17	Cool	3723
C18	Energetic, Powerful	158
C19	Soulful	369
C20	Hypnotic	216
C21	Sentimental	345
C22	Psychedelic	2335
C23	Lonely	111
C24	Spiritual	193
C25	Nostalgic	1800

6 RESULTS AND DISCUSSION

6.1 Classification Algorithms

Four renowned classification algorithms were attempted in this study, each of which has been extensively applied in former machine learning research work. Classification algorithms are supervised learning techniques, where a labelled dataset is required to be fed to them to obtain results.

Support Vector Machines (SVM): SVM [7] attempts to construct optimum hyperplanes which best separate a dataset into classes. A good separation is said to be obtained by finding the hyperplane which has largest distance to nearest training record from any class. Sequential Minimal Optimization (SMO) implementation of SVM was utilized in this research.

Naïve Bayes: Naïve Bayes [5] is based on Bayes' Theorem which describes the probability of an event, based on conditions that might be related to the event. Naïve Bayes algorithm is depicted by (5) where v_j depicts value of class attribute and a_i refers to value assumed by each of the other attributes. (There are j number of classes as $v_1, v_2, \ldots. v_j$).

$$h_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i|v_j) \quad (5)$$

Naïve Bayes assumption depicted by (6) is applied when using this algorithm.

$$P(a_1, a_2, \ldots\ldots, a_n|v_j) = \prod P(a_i|v_j) \quad (6)$$

C4.5: C4.5 [8] is a decision tree algorithm which is a variant of the original ID3 algorithm devised by Ross Quinlan. Information gain is the function used in this method to determine the attribute to be chosen at each level of the tree. Attribute with highest gain is chosen as root whereas the next attributes are chosen in descending order of their gains.

Random Forests: Random forests algorithm [6] forms a combination of prediction trees and produces the eventual result using ensemble methods. Thus the final class label is predicted via studying the class label prediction of each tree. Each tree is constructed using a sample of the training dataset, where each sample is created with replacement, adhering to bootstrapping mechanism. The overfitting problem of general decision trees is minimized by this classifier.

6.2 Experimental Setup

As mentioned before, merely the seven majority emotion classes were retained for further study. Since this dataset depicts class imbalance (number of data points from each class is not equal) to a certain extent, classification experiments were executed on undersampled and oversampled datasets as well. The undersampling algorithm applied, strives to reduce the number of data points from each class, to equal number of data points in class of smallest magnitude (C1). In oversampling mechanism suitable percentages must be provided for creation of synthetic data points in each class, to render magnitude of each class closer to magnitude of two largest classes; C7 and C17 (i.e. 200% adds another 2404 data points to C1 thus resulting in 3606 data points). Table 3 represents data distribution subsequent to undersampling and oversampling the dataset.

Table 3: Undersampling and oversampling the dataset

Emotion class	**No. of data points after undersampling**	**No. of data points after oversampling**
C1	1202	3606
C2	1202	3468
C3	1202	3278
C7	1202	3821
C17	1202	3723
C22	1202	3502
C25	1202	3600

The evaluation metrics utilized to assess aptitude of classifiers are as follows.

(The abbreviations stand for; TP – True Positive, TN – True Negative, FP – False Positive and FN – False Negative.)

Recall: Also known as true positive rate, sensitivity and hit rate, this helps evaluate the probability of correctly labeling members of the target class [13].

Recall = TP/ (TP + FN).

Precision: Also known as positive predictive value, this helps evaluate the probability that a positive prediction is correct [13].

Precision = TP/ (TP + FP).

F-measure: This is the harmonic mean of precision and recall [13].

F-measure =

2 * Precision * Recall/ (Precision + Recall).

False Positive Rate: Otherwise known as false alarm rate, this reflects the probability of falsely rejecting the null hypothesis for a particular test (i.e. classifier inaccurately states that an instance belonging to negative class is positive) [13].

False Positive Rate = FP/ (FP + TN)

Accuracy: This measure evaluates the proportion of correct predictions by a classifier.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Area under the Curve: AUC [13] is a graphical measure which depicts the area under Receiver operating characteristic (ROC) curve. ROC curve represents true positive rate against false positive rate, thus providing with an indication regarding overall performance of classifier. Table 4 depicts classifier evaluation based on AUC value.

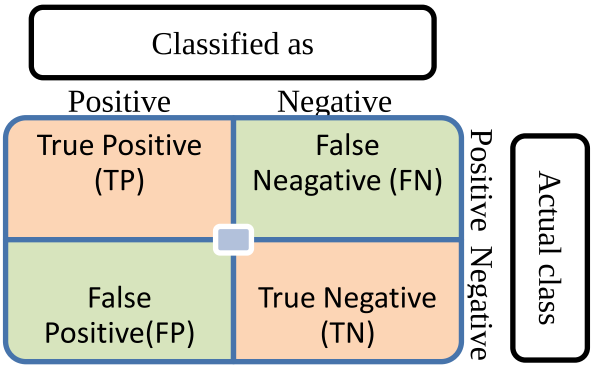
Figure 2 depicts the confusion matrix on which each evaluation measure is based. The two classes concerned are the 'positive' class and the 'negative' class. Classification output is labelled as TP, TN, FP and FN by considering the actual class of a data point and the class into which it is classified. Obtaining relatively high values for Recall, Precision, F-measure, Accuracy and AUC is preferred, whereas acquiring low values for False Positive Rate is desired.

6.3 Audio Based Classification Experiment

Figure 3 depicts accuracy of audio based classification. Thus, Random Forest with oversampling could be inferred as the best solution, based on Accuracy metric.

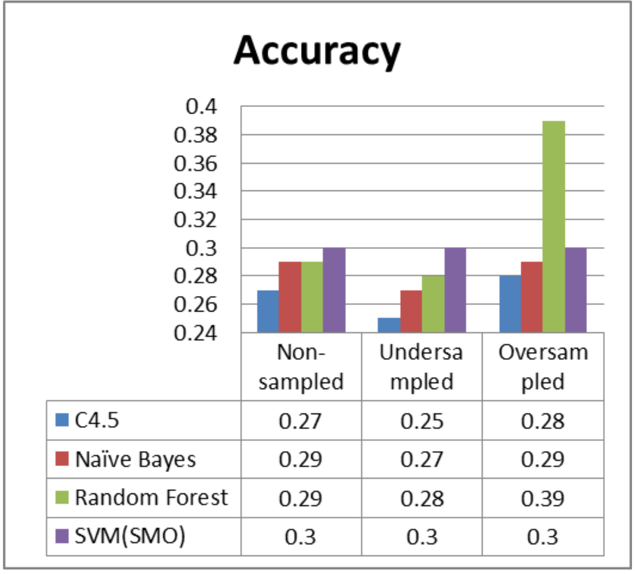
Table 4: Classifier performance based on AUC value

Area under the Curve	Reflection on performance
0.9 - 1	Excellent
0.8 – 0.9	Good
0.7 – 0.8	Fair
0.6 – 0.7	Poor
0.5 – 0.6	Fail



	Classified as: Positive	Classified as: Negative	Actual class
	True Positive (TP)	False Neagative (FN)	Positive
	False Positive(FP)	True Negative (TN)	Negative

Figure 2: Confusion matrix



	Non-sampled	Undersampled	Oversampled
C4.5	0.27	0.25	0.28
Naïve Bayes	0.29	0.27	0.29
Random Forest	0.29	0.28	0.39
SVM(SMO)	0.3	0.3	0.3

Figure 3: Accuracy of audio based classification

Table 5: Legend for figures 4, 5, 6, 7 and 8

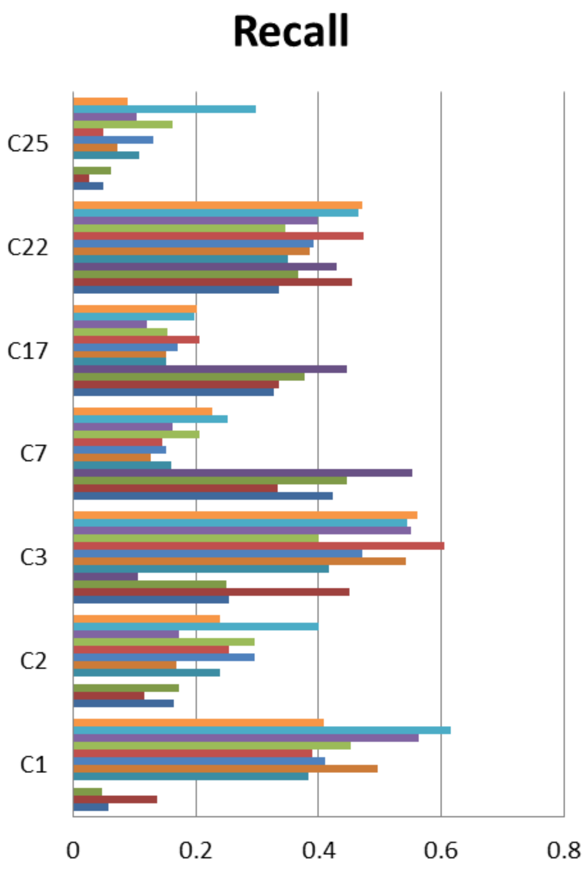
OS → Oversample + SVM
OR → Oversample + Random Forest
ON → Oversample + Naïve Bayes
OC → Oversample + C4.5
US → Undersample + SVM
UR → Undersample + Random Forest
UN → Undersample + Naïve Bayes
UC → Undersample + C4.5
NS → Non-sample + SVM
NR → Non-sample + Random Forest
NN → Non-sample + Naïve Bayes
NC → Non-sample + C4.5

Legend related to Figures 4 to 13 is depicted in Table 5.

According to Figure 4 which depicts values obtained for recall metric, classes C1, C2 and C25 have achieved the best results with Oversampling + Random Forest. C3 and C22 have obtained best results with Undersampling + SVM whereas classes C7 and C17 have obtained best results with Non-sampled SVM.

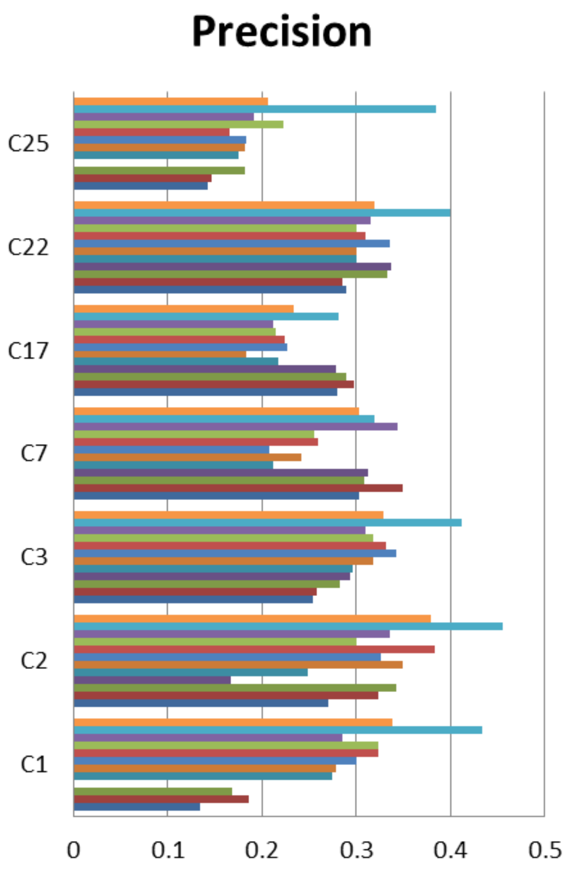
Figure 5 depicts results obtained for precision metric.

Best precision for classes C1, C2, C3, C22 and C25 has been obtained with Oversampling + Random Forest. Classes C7 and C17 however, have depicted best performance with Non-sampled Naïve Bayes.



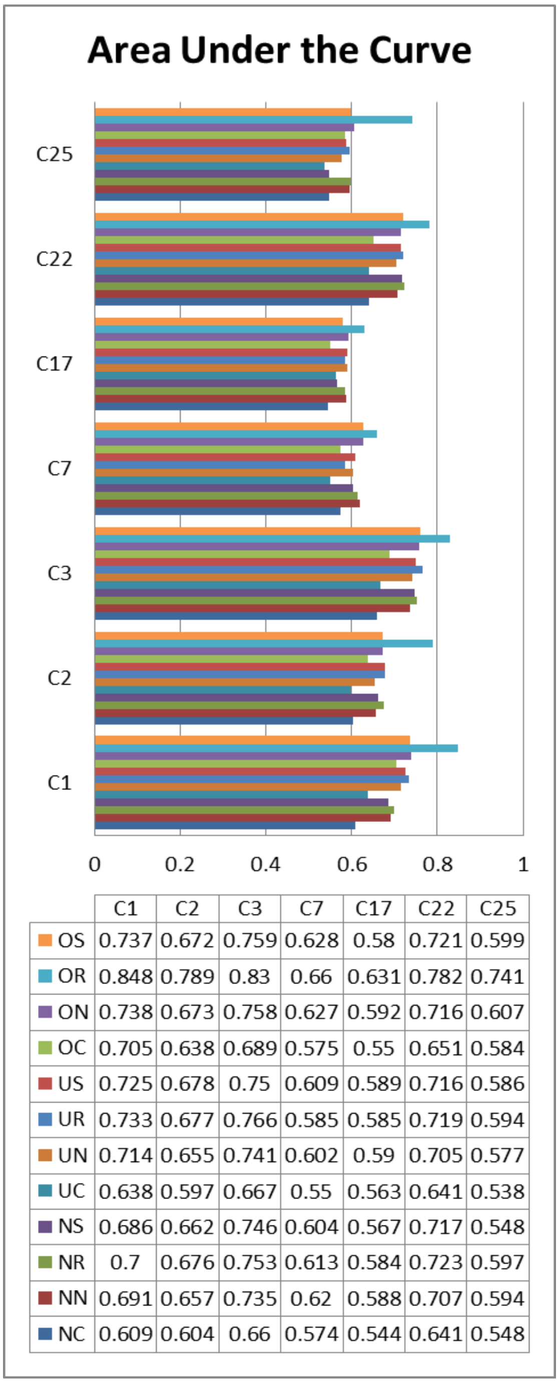
	C1	C2	C3	C7	C17	C22	C25
OS	0.409	0.24	0.561	0.227	0.203	0.471	0.089
OR	0.615	0.401	0.545	0.252	0.198	0.465	0.299
ON	0.563	0.173	0.551	0.162	0.12	0.399	0.103
OC	0.453	0.295	0.401	0.207	0.154	0.346	0.163
US	0.389	0.255	0.606	0.146	0.207	0.473	0.049
UR	0.411	0.295	0.471	0.151	0.171	0.391	0.131
UN	0.496	0.169	0.542	0.127	0.152	0.386	0.072
UC	0.384	0.239	0.418	0.16	0.152	0.35	0.107
NS	0	0.001	0.106	0.552	0.446	0.43	0
NR	0.047	0.173	0.251	0.446	0.377	0.367	0.063
NN	0.137	0.117	0.451	0.334	0.335	0.454	0.026
NC	0.058	0.165	0.254	0.424	0.328	0.335	0.049

Figure 4: Recall: Audio based classification



	C1	C2	C3	C7	C17	C22	C25
OS	0.338	0.379	0.329	0.303	0.233	0.319	0.206
OR	0.434	0.455	0.412	0.32	0.281	0.4	0.385
ON	0.285	0.336	0.31	0.344	0.212	0.316	0.192
OC	0.324	0.3	0.318	0.255	0.215	0.3	0.223
US	0.323	0.383	0.332	0.26	0.224	0.31	0.166
UR	0.3	0.326	0.343	0.208	0.227	0.336	0.184
UN	0.278	0.349	0.318	0.242	0.184	0.301	0.182
UC	0.275	0.249	0.296	0.212	0.218	0.301	0.175
NS	0	0.167	0.293	0.313	0.279	0.337	0
NR	0.169	0.342	0.282	0.308	0.289	0.333	0.182
NN	0.186	0.324	0.258	0.349	0.297	0.286	0.147
NC	0.135	0.271	0.254	0.303	0.28	0.29	0.142

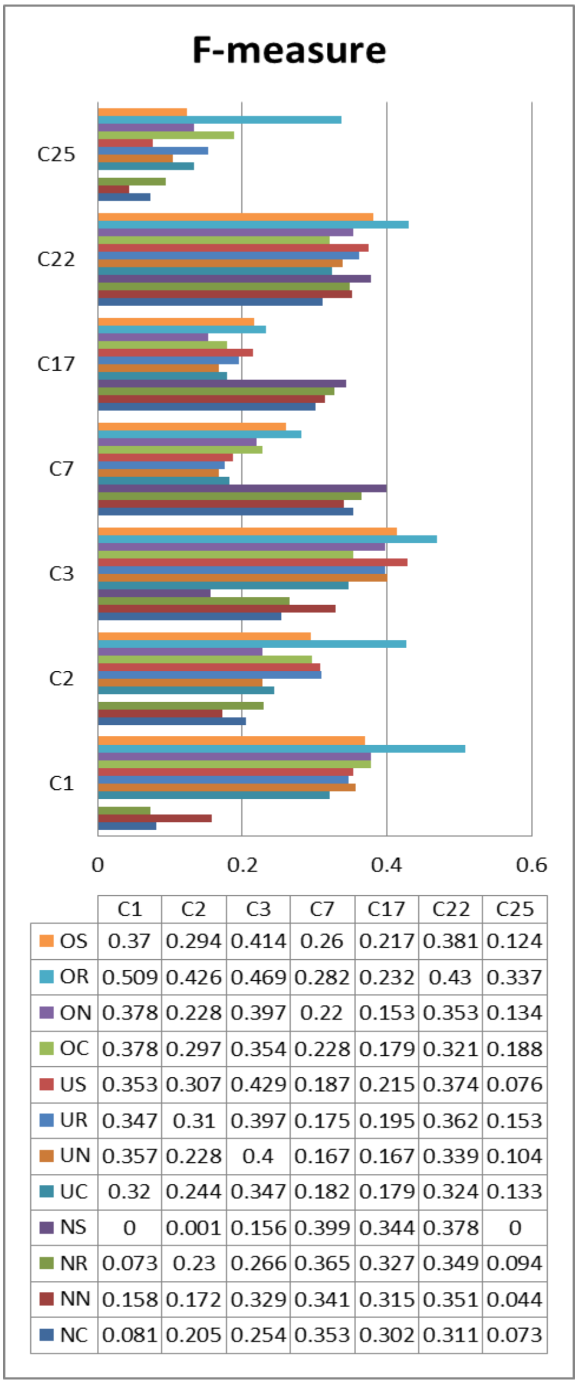
Figure 5: Precision: Audio based classification



	C1	C2	C3	C7	C17	C22	C25
OS	0.737	0.672	0.759	0.628	0.58	0.721	0.599
OR	0.848	0.789	0.83	0.66	0.631	0.782	0.741
ON	0.738	0.673	0.758	0.627	0.592	0.716	0.607
OC	0.705	0.638	0.689	0.575	0.55	0.651	0.584
US	0.725	0.678	0.75	0.609	0.589	0.716	0.586
UR	0.733	0.677	0.766	0.585	0.585	0.719	0.594
UN	0.714	0.655	0.741	0.602	0.59	0.705	0.577
UC	0.638	0.597	0.667	0.55	0.563	0.641	0.538
NS	0.686	0.662	0.746	0.604	0.567	0.717	0.548
NR	0.7	0.676	0.753	0.613	0.584	0.723	0.597
NN	0.691	0.657	0.735	0.62	0.588	0.707	0.594
NC	0.609	0.604	0.66	0.574	0.544	0.641	0.548

Figure 6: Area under the curve (AUC): Audio based classification

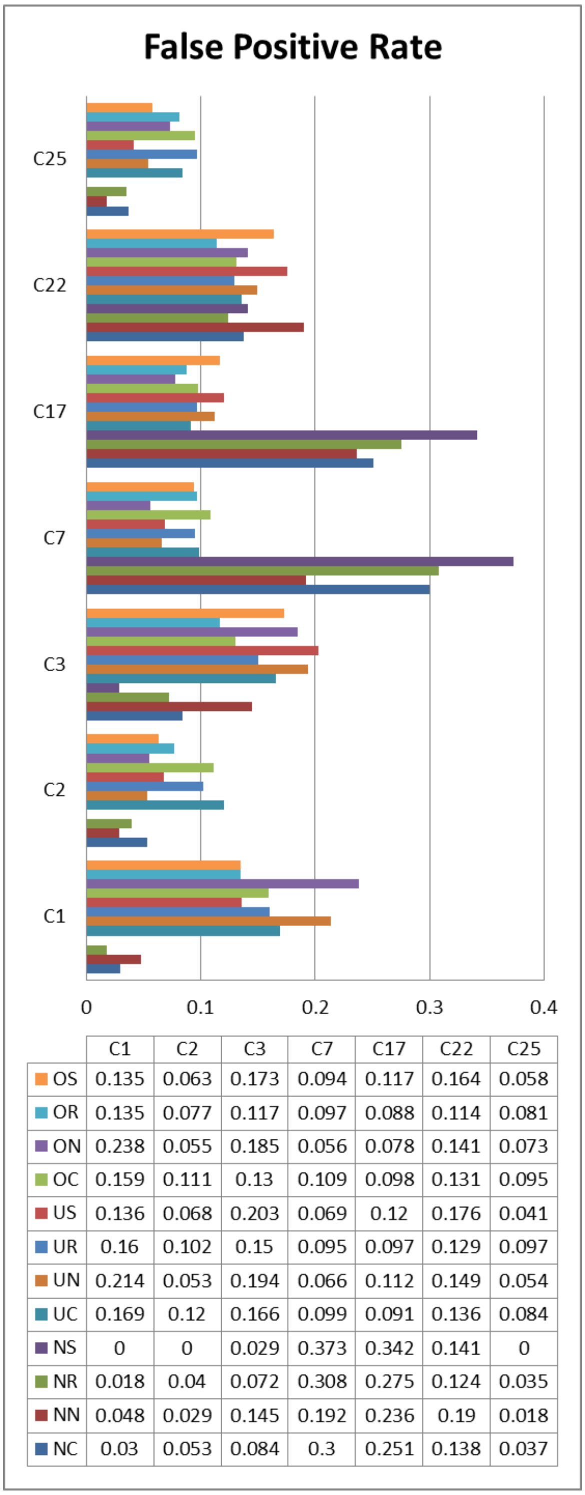
Figure 6 depicts values obtained for area under the curve measure (AUC). Best results were obtained for each class when Random Forest algorithm was executed on Oversampled dataset. According to Table 4 classification performance of Random Forest with Oversampling could be categorized as ‘good’ for classes C1 and C3, whereas it has performed ‘fairly’ for classes C2, C22 and C25.



	C1	C2	C3	C7	C17	C22	C25
OS	0.37	0.294	0.414	0.26	0.217	0.381	0.124
OR	0.509	0.426	0.469	0.282	0.232	0.43	0.337
ON	0.378	0.228	0.397	0.22	0.153	0.353	0.134
OC	0.378	0.297	0.354	0.228	0.179	0.321	0.188
US	0.353	0.307	0.429	0.187	0.215	0.374	0.076
UR	0.347	0.31	0.397	0.175	0.195	0.362	0.153
UN	0.357	0.228	0.4	0.167	0.167	0.339	0.104
UC	0.32	0.244	0.347	0.182	0.179	0.324	0.133
NS	0	0.001	0.156	0.399	0.344	0.378	0
NR	0.073	0.23	0.266	0.365	0.327	0.349	0.094
NN	0.158	0.172	0.329	0.341	0.315	0.351	0.044
NC	0.081	0.205	0.254	0.353	0.302	0.311	0.073

Figure 7: F-measure: Audio based classification

Figure 7 depicts results obtained for f-measure. Classes C1, C2, C3, C22 and C25 have obtained best values for f-measure with Oversampling + Random Forest. Classes C7 and C17 have achieved best results with Non-sampled SVM.



	C1	C2	C3	C7	C17	C22	C25
🟧 OS	0.135	0.063	0.173	0.094	0.117	0.164	0.058
🟦 OR	0.135	0.077	0.117	0.097	0.088	0.114	0.081
🟪 ON	0.238	0.055	0.185	0.056	0.078	0.141	0.073
🟩 OC	0.159	0.111	0.13	0.109	0.098	0.131	0.095
🟥 US	0.136	0.068	0.203	0.069	0.12	0.176	0.041
🟦 UR	0.16	0.102	0.15	0.095	0.097	0.129	0.097
🟧 UN	0.214	0.053	0.194	0.066	0.112	0.149	0.054
🟦 UC	0.169	0.12	0.166	0.099	0.091	0.136	0.084
🟪 NS	0	0	0.029	0.373	0.342	0.141	0
🟩 NR	0.018	0.04	0.072	0.308	0.275	0.124	0.035
🟥 NN	0.048	0.029	0.145	0.192	0.236	0.19	0.018
🟦 NC	0.03	0.053	0.084	0.3	0.251	0.138	0.037

Figure 8: False positive rate: Audio based classification

Unlike for other metrics, obtaining minimal value for false positive rate is preferred. Hence, as shown in Figure 8 which depicts values obtained for false positive rate, classes C1, C2, C3 and C25 have achieved best results with Non-sampled SVM method. Classes C7 and C17 have performed best with Oversampling + Naïve Bayes. C22 class has obtained the best classification result with Oversampling + Random Forest.

Table 6: Best classification algorithm according to different metrics: Audio based classification

	Recall	**Precision**	**AUC**	**F-measure**	**FP rate**
C1	OR	OR	OR	OR	NS
C2	OR	OR	OR	OR	NS
C3	US	OR	OR	OR	NS
C7	NS	NN	OR	NS	ON
C17	NS	NN	OR	NS	ON
C22	US	OR	OR	OR	OR
C25	OR	OR	OR	OR	NS

Table 6 summarizes the results obtained for each metric in audio based classification. Random Forest with Oversampling has qualified as the best classifier in 21 out of 35 instances according to Table 6 (when metric + class pairs are considered individually, 35 instances are obtained). Thus it could be assigned an aptitude score of 0.6 (21/35 = 0.6). Scores assigned for other algorithms are as follows.

Non-sampled SVM = 0.22
Undersampled SVM = 0.06
Non-sampled Naïve Bayes = 0.06
Oversampled Naïve Bayes = 0.06

Random Forest with Oversampling could be assigned a higher score when considering the algorithm which has produced best results with respect to different metric and individual classes. Furthermore, in the instances where Oversampling + Random Forest hasn't qualified as the best classifier, it's often included among the best five. Figure 3 depicts that best accuracy could be obtained with Oversampling + Random Forest. Thus, for audio based classification, Random Forest with Oversampling has qualified as the comparative, best classifier.

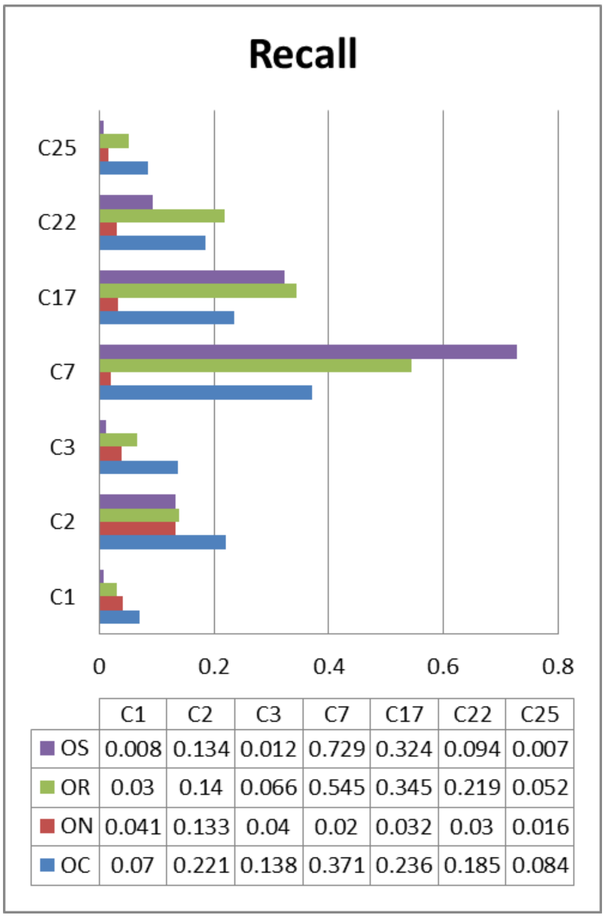
Due to the positive impact oversampling depics on the experiments, lyric based classification was executed on the oversampled dataset alone.

6.4 Lyric Based Classification Experiment

As depicted by Figure 9, classes C1, C2, C3 and C25 have obtained best classification accuracy with C4.5 algorithm. Random Forest has produced best results for C22 and C17 whereas SVM has surpassed other algorithms for class C7.

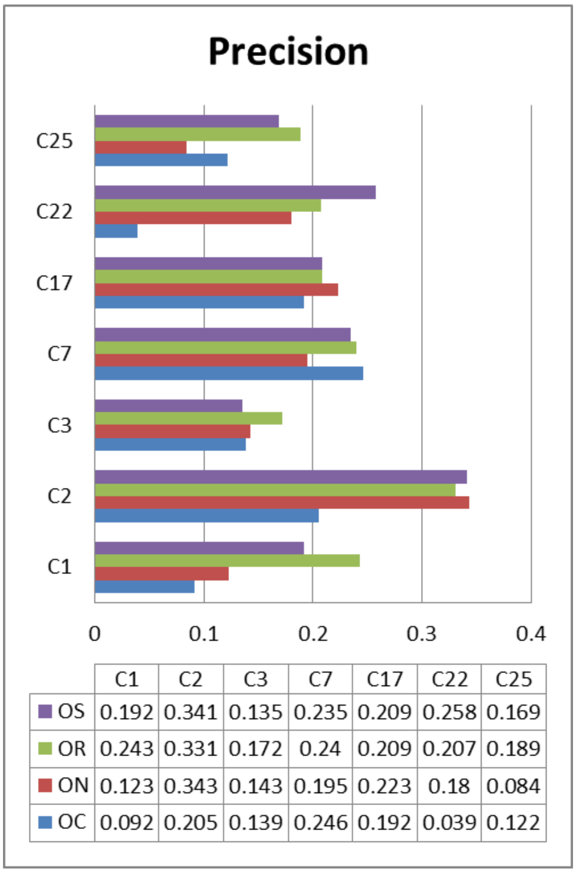
Figure 10 depicts the precision acquired with lyric based classification. Random Forest has performed best for classes C1, C3 and C25 according to precision metric whereas Naïve Bayes has produced best results for classes C2 and C17. C4.5 has given best results for C7 and SVM for C22.

According to AUC metric, Random Forest has outperformed the rest for all classes as shown in Figure 11



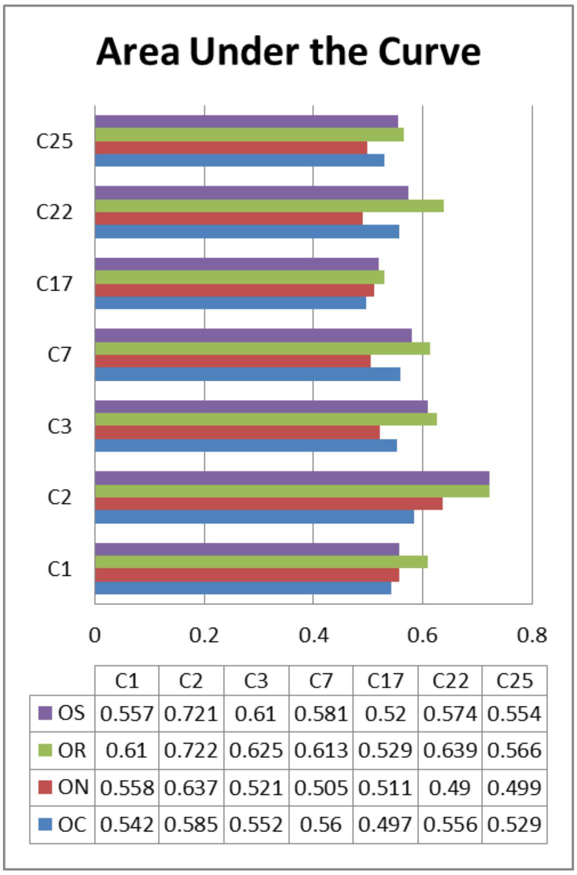
	C1	C2	C3	C7	C17	C22	C25
🟪 OS	0.008	0.134	0.012	0.729	0.324	0.094	0.007
🟩 OR	0.03	0.14	0.066	0.545	0.345	0.219	0.052
🟥 ON	0.041	0.133	0.04	0.02	0.032	0.03	0.016
🟦 OC	0.07	0.221	0.138	0.371	0.236	0.185	0.084

Figure 9: Recall: Lyric based classification



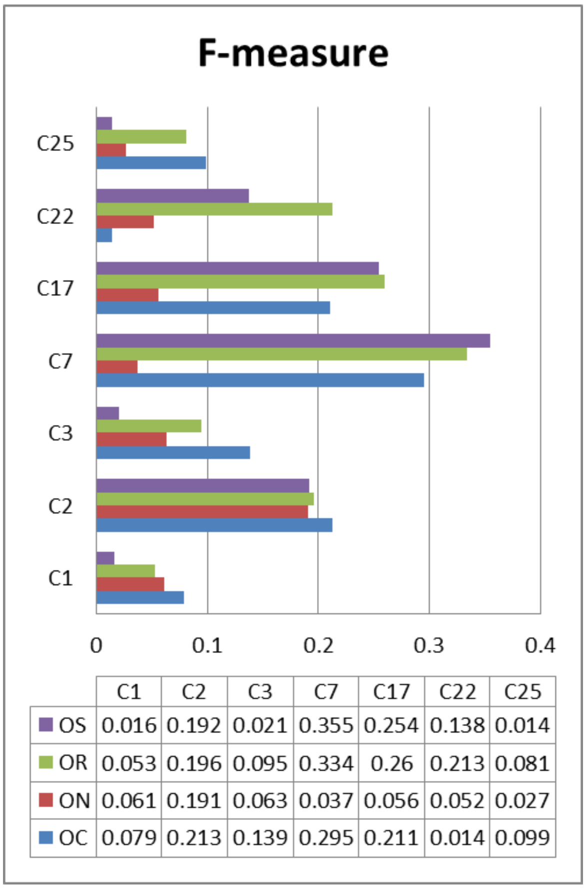
	C1	C2	C3	C7	C17	C22	C25
🟪 OS	0.192	0.341	0.135	0.235	0.209	0.258	0.169
🟩 OR	0.243	0.331	0.172	0.24	0.209	0.207	0.189
🟥 ON	0.123	0.343	0.143	0.195	0.223	0.18	0.084
🟦 OC	0.092	0.205	0.139	0.246	0.192	0.039	0.122

Figure 10: Precision: Lyric based classification



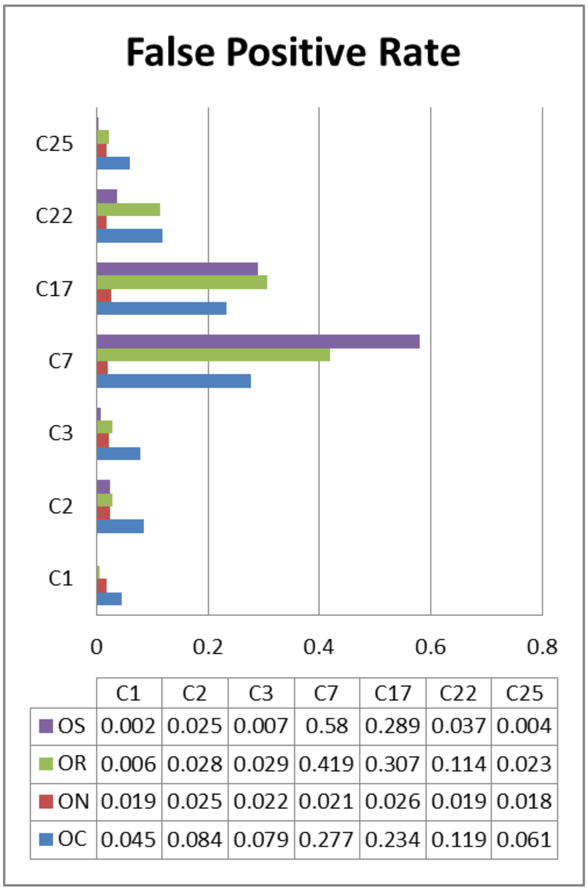
	C1	C2	C3	C7	C17	C22	C25
🟪 OS	0.557	0.721	0.61	0.581	0.52	0.574	0.554
🟩 OR	0.61	0.722	0.625	0.613	0.529	0.639	0.566
🟥 ON	0.558	0.637	0.521	0.505	0.511	0.49	0.499
🟦 OC	0.542	0.585	0.552	0.56	0.497	0.556	0.529

Figure 11: AUC: Lyric based classification



	C1	C2	C3	C7	C17	C22	C25
🟪 OS	0.016	0.192	0.021	0.355	0.254	0.138	0.014
🟩 OR	0.053	0.196	0.095	0.334	0.26	0.213	0.081
🟥 ON	0.061	0.191	0.063	0.037	0.056	0.052	0.027
🟦 OC	0.079	0.213	0.139	0.295	0.211	0.014	0.099

Figure 12: F-measure: Lyric based classification



	C1	C2	C3	C7	C17	C22	C25
🟪 OS	0.002	0.025	0.007	0.58	0.289	0.037	0.004
🟩 OR	0.006	0.028	0.029	0.419	0.307	0.114	0.023
🟥 ON	0.019	0.025	0.022	0.021	0.026	0.019	0.018
🟦 OC	0.045	0.084	0.079	0.277	0.234	0.119	0.061

Figure 13: FP-rate: Lyric based classification

Figure 12 is a graphical representation of how well classifiers have performed with respect to F-measure metric. According to that, C17 and C22 have shown best results for Random Forest. C1, C2, C3 and C25 have depicted best results for C4.5 whereas C7 has shown best performance for SVM.

As mentioned before, getting minimal value for False Positive Rate is desired. Thus, according to Figure13, classes C1, C2, C3 and C25 have shown best performance with SVM. C7, C17 and C22 have performed best with Naïve Bayes.

Table 7: Best classification algorithm according to different metrics: Lyric based classification

	Recall	**Precision**	**AUC**	**F-measure**	**FP rate**
C1	OC	OR	OR	OC	OS
C2	OC	ON	OR	OC	OS
C3	OC	OR	OR	OC	OS
C7	OS	OC	OR	OS	ON
C17	OR	ON	OR	OR	ON
C22	OR	OS	OR	OR	ON
C25	OC	OR	OR	OC	OS

Table 7 summarizes the performance analysis of lyric based classification. Only the oversampled dataset was used for lyric based classification. An aptitude score could be assigned to each lyric based classification experiment as follows.

C4.5 = 0.257
SVM = 0.2
Random Forest = 0.4
Naïve Bayes = 0.143

Random Forest has qualified as the best classifier for lyric based classification as well, as conveyed by the associated aptitude scores. Furthermore, the AUC score depicts that Random Forest outperforms any other classifier for lyric based classification.
Oversampling is a technique which ensures that valuable information in the dataset is not lost when attempting to balance the dataset. Rather, it creates synthetic data in emotion classes with fewer data points, such that the characteristics of the newly created data are

similar to other data points in the respective class. On the contrary, undersampling results in loss of certain information, which is a probable reason why oversampling performs better than undersampling.

Random forest is an ensemble method, where a multitude of decision trees are created when creating classification model. This allows selecting the 'most probable' emotion class of a given song. The other three classification methods attempted however, do not adopt this ensemble method. Rather, they classify a song to a single specific emotion class alone, due to those being devoid of an interim phase where several probable classes are found. Since a given song often has the potential of evoking several emotions, Random forest seemingly has performed better at predicting the 'most likely' emotion class to which a song belongs.

7 CONCLUSION AND FUTURE WORK

Potential for music to evoke emotion in individuals has forever been an inexplicable, yet evident phenomenon. While numerous research work has been conducted to deduce whether the incurrence of emotion when listening to music is scientifically explicable, the literature survey conducted helped realize that there existed numerous limitations with existent mechanisms. One such limitation was usage of generic emotion models for predicting emotions in music. It has been observed that emotions which are incurred when listening to music take a different form than generic emotions. For instance, the sadness evoked by music is not necessarily as intense as sadness evoked from real life experience. In fact, sadness (melancholic) belonged to the same cluster as calming and romantic in the music specific emotion model created based on the MSD dataset. Thus creation of a music specific emotion model for classification of music was viewed to be a positive aspect of this study. Seven significant emotion classes were thus identified, which are depicted in Table 8.

Table 8: Music emotion classes identified with emotion model

Musically induced emotion
Joyful, Danceable
Witty
Calming, Melancholic, Romantic
Sensual
Cool
Psychedelic
Nostalgic

A series of experiments were conducted to attempt classification of music into recognized emotion classes. Of the classification attempts, the best music emotion prediction model was provided as a combination of Random Forest with oversampling. This fact is supported by the aptitude scores provided in Table 9, which summarizes the performance of each classifier for audio and lyric based classification.

Table 9: Summary of results

Classifier	Classification method			
	Audio		Lyric	
	Sampling	Score	Sampling	Score
C4.5	-	-	OC	0.257
SVM	NS	0.22	OS	0.2
	US	0.06		
Random Forest	OR	0.6	OR	0.4
Naïve Bayes	NN	0.06	ON	0.143
	ON	0.06		

A simple comparison of the AUC values helps deduce the fact that audio based classification is superior to lyric based classification. Fair performance has only been obtained for C2 in lyric based classification, according to Table 4. Thus it could be concluded that audio features are more apt for forming music emotion recognition models.

The emotion model formed in this study was reliant on the assumption that emotion related tags associated with music convey perceived emotion. This assumption may not always be valid since words such as ‘love’ could be conveying that the song is a love song, but not that the emotion was felt by a person. This limitation could be resolved by collecting information from a number of people regarding what emotions they experienced when listening to songs from a dataset.

REFERENCES

[1] Princeton University, “About WordNet. WordNet,” 2010. [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 02-Aug-2015].

[2] A. Gabrielsson, “Emotion perceived and emotion felt: Same or different?,” *Music. Sci.*, vol. 5, no. 1, pp. 123–147, 2002.

[3] K. R. Scherer and M. R. Zentner, “Emotional effects of music: production rules,” in *Music and Emotion: Theory and Research*, 2001, pp. 361–387.

[4] M. Zentner, D. Grandjean, and K. R. Scherer, “Emotions evoked by the sound of music: characterization, classification, and measurement.,” *Emotion*, vol. 8, no. 4, pp. 494–521, Aug. 2008.

[5] I. Rish, “An empirical study of the naive Bayes classifier,” *Int. Jt. Conf. Artif. Intell.*, vol. 3, no. 22, pp. 41–46, 2001.

[6] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[7] I. Steinwart and A. Christmann, *Support Vector Machines*. New York: Springer, 2008, pp. 1–25.

[8] N. V Chawla, “C4 . 5 and Imbalanced Data sets : Investigating the effect of sampling method , probabilistic estimate , and decision tree structure,” in *Proc. Int’l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II*, 2003, p. 8.

[9] P. Kanters, “Automatic mood classification for music,” Tilburg University, Netherlands, 2009.

[10] J. Han and M. Kamber, *Data mining: concepts and techniques*, 2nd ed. Elsevier, 2006.

[11] S. Guha, R. Rastogi, and K. Shim, “ROCK: a robust clustering algorithm for categorical attributes,” in *Proceedings., 15th International Conference on Data Engineering*, 1999, pp. 512 – 521.

[12] T. Hill and P. Lewicki, *Statistics: Methods and Applications*, 2nd ed. StatSoft, Inc., 2007, p. 800.

[13] D. M. . Powers, “Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.