# Machine Learning Classification Models

## Logistic regression

Logistic regression (LR) is a supervised learning algorithm that can be used to solve classification problems. LR is one of the most popular methods that have been used in classification tasks (Ajoodha, et al., 2015) which used logistic function to predict the categorical dependent variable. Where x represents the features, g(z) is the activation function and h(x) is the hypothesis function. Then we can illustrate the definition of the logistic model as:

### Advantages

* It performs well even when the dataset size is small.
* Because logistic regression calculates coefficients for each factor, it is simpler to understand how each feature influences the result.
* It has a low resource requirement and is computationally effective.
* Provides probabilities for every category that can be used to gauge the model's level of competence.

### Disadvantages

* Considers a linear decision boundary, therefore it is inappropriate for interactions that are complex and non-linear.
* Multicollinearity: If features are significantly linked, efficiency decreases.
* Robust to outliers that could have a negative impact on performance.
* Limited Complexity: For sophisticated data sets with intricate interactions between features, less effective than other methods.

## Naïve Bayes

Naïve Bayes algorithm (NB) is a supervised learning algorithm, depending on Bayes theorem which is used for solving classification problems like text classification (Bužić & Dobša, 2018) with a high-dimensional training dataset. Naïve Bayes algorithm helps us to build a fast machine learning model that can make quick predictions depending on the probability of an object.

Naive Bayes predicts depending on past experience. Below mentioned equation determined the working of NB.

Where:

* *X: Unknown class’s data*
* *c: Specific class*
* *P(c | x): posterior probability*
* *P (c): prior probability*
* *P(X | c): Probability that is based on conditions on the hypothesis*
* *P (x): Probability A*

### Advantages

### It is simple and straightforward to grasp because it is built on the Bayes theorem and implies prediction independence.

### Because of its simplicity, it is able to deal with big datasets efficiently. Computationally efficient, it takes lesser time to train large dataset.

### Easily tackles challenges with multiclass classification.

### Excellent in text data classification.

### Disadvantages

* It have naive assumption that all features are independent but it’s not possible in real-time dataset.
* Probability estimations are frequently unreliable, despite the fact that classification may be precise.
* The algorithm gives a class zero probability when it does not appear in the training data
* It is a simple algorithm and not appropriate for really complex data.

## Random Forest Classifier

A Random Forest Classifier is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions using a majority voting mechanism. For multi-class problems (Bahuleyan, 2018), each decision tree in the random forest is trained on a bootstrapped sample of the dataset and uses a random subset of features at each split. This strategy introduces diversity among the trees, reducing overfitting and improving generalization. The predicted class for an instance is the one with the majority vote among all trees. Random forests are robust to overfitting, handle non-linear relationships well, and often achieve better performance compared to individual decision trees.

### Advantages

### Proficient of detecting associations among features that are not linear.

### It reduces probability of overfitting by averaging several decision trees,

### It specifies information about the significance of a feature, assisting in feature selection.

### Possibility of employing proximity-weighted imputation or median values to address missing data.

### Disadvantages

* Random forests can be complicated, requiring more memory and processing power.
* Due to the construction of numerous trees, this model requires more time to train than simpler models.
* Compared to logistic regression, it is a black-box model that is more difficult to comprehend.
* When dealing with noisy data, it may be possible to biased to the noise.

# LDA Topic Model

A particular type of text mining is called topic modelling which extracts recurring patterns from text-based information. It involves applying models to the collection of text data in order to create a representation of the subjects covered there. An NLP method known as topic modelling uses statistical connections of words in text data to construct latent topics. Topic modelling is a method of information extraction and classification of enormous collections of texts. Without using any predetermined dictionaries or interpretive rules, these latent themes are constructed. A set of texts' fundamental subjects are discovered using topic modelling, which also provides quantitative measurements that identify and characterise the topics of specific texts among a set of texts and enable thematic comparisons both inside and between texts. Gensim library has been used to build model. To build LDA on the dataset we have extracted unigram features from the text. We have bag of words (Bow) features from the dataset and apply LDA model on extracted features. To chose optimal topic number for our dataset we have evaluated LDA model on topics numbers in range 5 to 30. To evaluate LDA model’s performance, we have used the Coherence evaluation metric the highest score determines the best number of topics. LDA model extracts 7 number of topics over the entire dataset having coherence score 0.401.

# Experimental Setup

This segment designates the experimental setup used for developing machine learning techniques for genre and topic prediction tasks and topic modeling using music lyrics dataset. This module illustrates experimental settings and evaluation methods.

## Evaluation Methodology

The problems of genre prediction and topic prediction are treated as supervised multiclass classification (Prajwal, et al., 2021) tasks. For genre prediction task input will be lyrics and other supporting features according to research question and output attribute will be genre. For topic prediction task input will be lyrics and other supporting features according to research question and output attribute will be topic. The problem of theme extraction from text is treated as topic modeling task.

## Models

For the multiclass classification task we have built Logistic Regression, Naïve Bayes and Random Forest classifier. We have extracted features using tfidf vectorizer. As it can be seen from above exploratory data analysis dataset is highly imbalanced to balance the dataset we have used SMOTE oversampling technique. Synthetic Minority Over-sampling Technique (SMOTE) is an advanced oversampling method that generates synthetic samples for the minority class to balance the class distribution. SMOTE works by selecting samples from the minority class and generating new synthetic instances based on the feature space similarities between the selected samples and their k-nearest neighbors. After applying SMOTE we passed those features to the models. After the model is trained on the dataset we have used the test data to evaluate models performance using multiple evaluation measures.

## Evaluation Measures

For experimentations evaluated in this project, we have used the following mentioned evaluation metrics. To compare different models, we used five metrics to evaluate the model’s performance. Starting with Accuracy, that is the most common criteria for measuring classification models performance, but if working with a multiclass classification dataset, this criterion is not appropriate because the minority class will have a small contribution to the accuracy criteria. So, we will use six evaluation measures including Accuracy, Recall, Precision, F1 Score and Confusion Matrix. The formulas are defined as follows:

**Accuracy**

Accuracy provides the ratio among true predictions and overall predictions of the model.

**Recall**

It depicts that how many classes from positive class are actually model classified as positive.

**Precision**

It depicts that how many classes predicted by model as positive are actually positive.

**F1 score**

It provides the average score between precision and recall.

**Confusion Matrix**

It provides the complete frequency of TP, FN, TN, and FP in overall model predictions.

Where TP, FN, TN, and FP given in equations represent the true positive prediction, false negative prediction, true negative prediction, false positive prediction, respectively.

# Libraries and Toolkits

We executed machine learning classification experiments using Scikit learn machine learning library. To build LDA topics model we have used genism library. All the experiments are performed in Python language. We have performed all experiments in Jupyter Notebook with system specifications, processor core i7, and RAM 8GB.

**RQ 1:** How accurately can machine learning models classify songs based on their lyrics and topics?

To address this research question, we have combined the lyrics and predefined topics attributes after extracting features from both attributes we pass the features to machine learning models. Among the algorithms analyzed, Random Forest emerged as the most proficient model, achieving an accuracy of 71.09%, a recall of 71.09%, a precision of 71.32%, and an F1-Score of 70.98%. In contrast, Logistic Regression demonstrated moderate performance with an accuracy of 58.81%, and Naïve Bayes was the least effective with an accuracy of 51.35%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 58.81 | 58.81 | 57.75 | 58.12 |
| Naïve Bayes | 51.35 | 51.35 | 50.2 | 49.57 |
| Random Forest | 71.09 | 71.09 | 71.32 | 70.98 |

Table 1: Results of genre prediction using lyrics and topics

These findings offer valuable insights into the capabilities and limitations of different machine learning algorithms for the prediction of genre based on lyrics and topics.

**RQ 2:** How do lyrical topics vary across different genres?

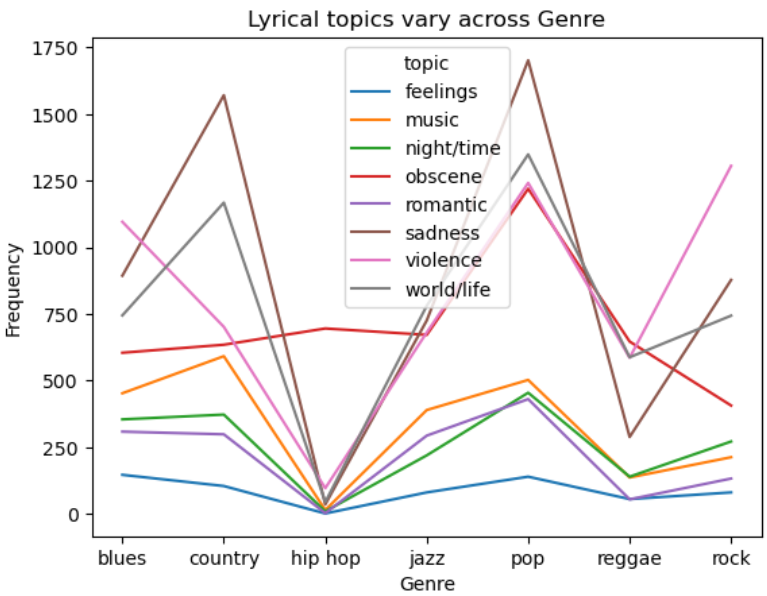


Figure 4: Lyrical topics vary across different genres

Feelings topic manifests a relatively uniform distribution across all genre lyrics with a slight augmentation observed particularly within the genres of blues, pop and rock. Music topic is predominantly features in country and pop genres, in the same manner we plotted all other features frequency in Fig 4. Higher number of topic themes belongs to country and pop genre. Topics frequency in specific genre also vary with the count of specific genre instances in the dataset.

**RQ 3:** Can a song's release decade be accurately predicted from its lyrics and topics?

We combined the lyrics and topics attributes and extracted features in combination and pass that data to models. Among the algorithms analyzed, Random Forest emerged as the most proficient model, achieving an accuracy of 54.91%, a recall of 54.91%, a precision of 54.23%, and an F1-Score of 54.0%. In contrast, Naïve Bayes demonstrated moderate performance with an accuracy of 34.64%, and Logistic Regression was the least effective with an accuracy of 32.86%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 32.86 | 32.86 | 31.07 | 31.01 |
| Naïve Bayes | 34.64 | 34.64 | 35.43 | 32.87 |
| Random Forest | 54.91 | 54.91 | 54.23 | 54.0 |

Table 2: Results of release date prediction using lyrics and topics

It can be observed that machine learning models are not able to efficiently predict the release date based on song lyrics.

**RQ 4:** Can the song topic be predicted from the lyrics of a song?

It is discernible that machine learning models demonstrate adeptness in accurately forecasting the underlying topics by analyzing the textual content of song lyrics. Logistic Regression outperforms all other models having accuracy of 96.81%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 96.81 | 96.81 | 96.82 | 96.80 |
| Naïve Bayes | 92.56 | 92.56 | 92.64 | 92.50 |
| Random Forest | 90.42 | 90.42 | 90.29 | 90.32 |

Table 3: Results of topic prediction using lyrics

**RQ 5:** How accurately can a song's genre be predicted based on its lyrics alone, as compared to using lyrics in combination with topics?

Comparing the Table 1 and Table, there is a conspicuous enhancement in performance when exclusively employing song lyrics as input. Among the evaluated classifiers, the Random Forest algorithm emerges outstanding, achieving an accuracy of 70.42%, thus surpassing the efficacy of its counterparts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 58.81 | 58.81 | 57.71 | 58.11 |
| Naïve Bayes | 51.93 | 51.93 | 51.17 | 50.21 |
| Random Forest | 70.42 | 70.42 | 70.73 | 70.36 |

Table 4: Results of genre prediction using lyrics

**RQ 6:** What are the dominant themes in song lyrics over the entire dataset?

We have extracted 7 discriminant topics from the entire dataset. Based on keywords we assigned these 'music', 'obscene', 'romantic', 'violence', 'sadness', 'world/life', 'night/time' are the labels for all the themes extracted respectively. Below Fig 5. Shows the keywords wordcloud of each topic theme.



Figure 5: Keywords wordcloud of each topic theme

Fig 6. depicts the weightage of all topics extracted using LDA model from the entire dataset.

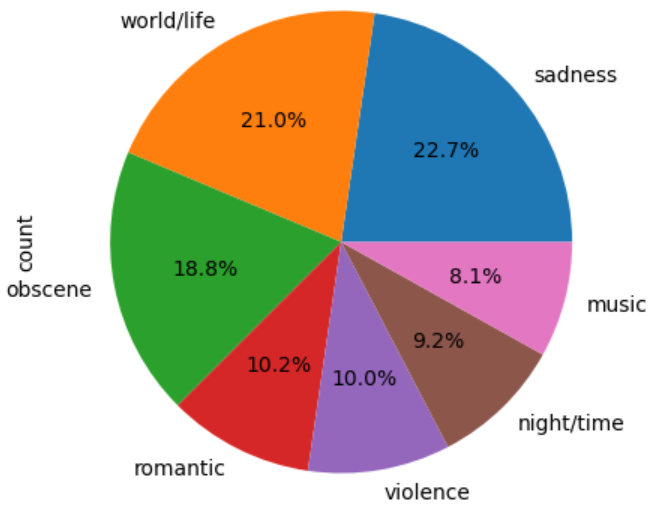


Figure 6: Weightage of all topics extracted using LDA model

**RQ 7:** How have topics in lyrics evolved over the years, from 1950 to 2019?

It can be observed from graph that trend of all topics increasing over the decades with minor fluctuations. Specifically, obscene topic has higher frequency than all others in year 2017 and 2018. After that Frequency of word/life and sadness topic is higher over the decades. In comparison to others frequency of romantic and music topics is lower.

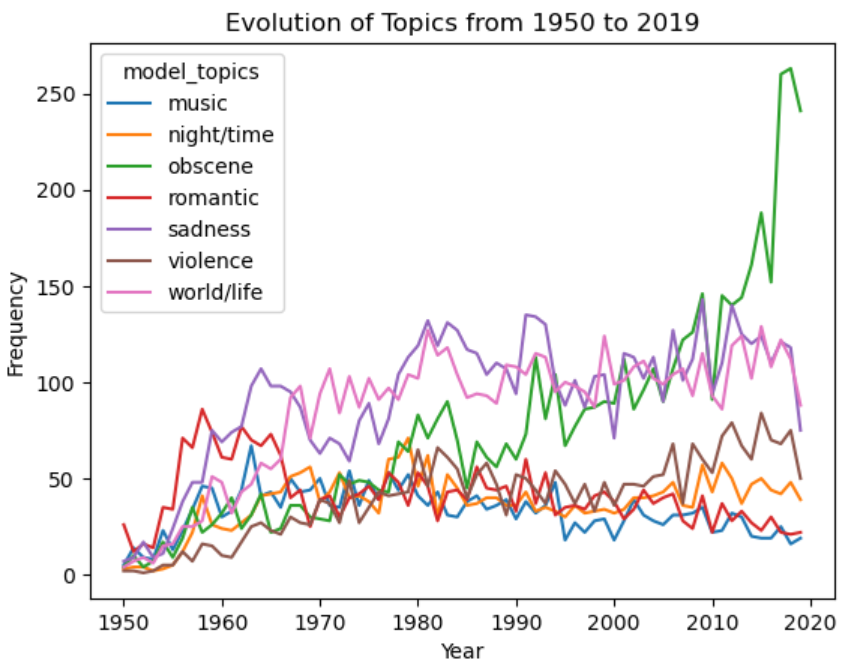


Figure 7: Evolution of Topics from 1950 to 2019

**RQ 8:** Are there distinctive topics that can be associated with different genres?

In the correlation graph, it can be observed that some points or clusters are more densely packed and situated along a line, either upward sloping that shows a positive correlation or downward sloping that shows a negative correlation. The strength of the correlation can often be interpreted through the density and slope of these points or clusters. From the graph it can be interpreted that there is not a strong correlation between genre and topics. Sadness topic have higher correlation with country and pop genre. After that sadness correlation with rock, obscene and word/life correlation with pop. Hip hop genre has lower correlation with all topics. Higher correlation shows that topics are most discussed in particular genre.

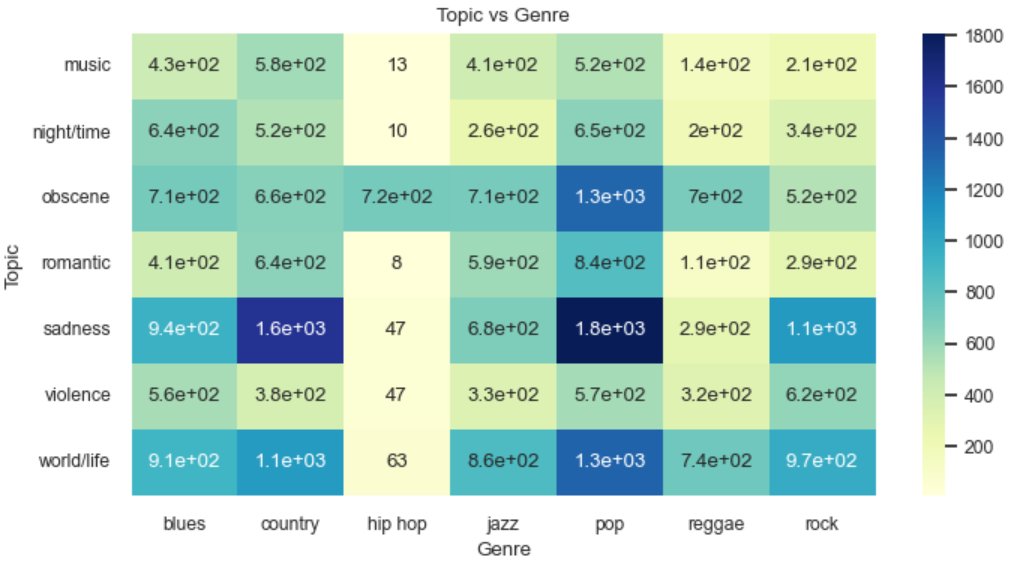


Figure 8: Topic vs Genre correlation

**RQ 9:** How do topics derived from the lyrics correlate with predefined topics?

It can be observed from graph that predefined topics in the dataset and topics extracted from LDA model are highly correlated that shows that our model was able to appropriately extract topics from the dataset that are most likely same as predefined topics for each lyric.

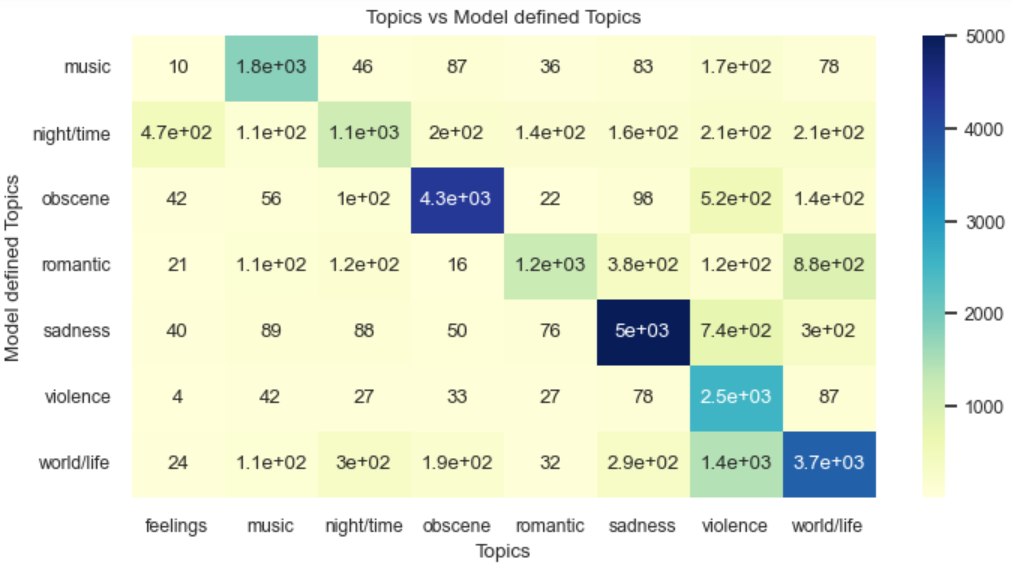


Figure 9: Topics vs Model defined Topics correlation

**RQ 10:** How do the topics present in lyrics relate to a song's release date or era?

After analyzing correlation between release date and topics it can be seen that there is no particular association in both attributes. Only obscene topic has higher correlation with year 2017, 2018 and 2019.

**RQ 11:** Can topic modeling keywords help improve the accuracy of genre prediction models when combined with the lyrics and predefined topics?

In this problem criteria Random Forest model performs better than all other algorithms having an accuracy of 59.56%. Comparing the Table 1 results with Table 5 shows that results decrease as we combined the lyrics and predefined topics with new extracted topic keywords. So we can say that, no topic modeling keywords did not help improve the accuracy of genre prediction models when combined with the lyrics and predefined topics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 59.25 | 59.25 | 58.59 | 58.81 |
| Naïve Bayes | 37.24 | 37.24 | 36.83 | 34.15 |
| Random Forest | 59.56 | 59.56 | 60.37 | 59.70 |

Table 5: Results of genre prediction using lyrics and topics and model extracted theme topics

**RQ 12:** Can topic modeling help improve the accuracy of predefined topic prediction models when combined with the lyrics?

Comparing the Table 1 results with Table 5 shows that results decrease as we combined the lyrics with new extracted topic keywords. So we can say that, no topic modeling did not help improve the accuracy of predefined topic prediction models when combined with the lyrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy (%)** | **Recall (%)** | **Precision (%)** | **F1 (%)** |
| Logistic Regression | 93.24 | 93.24 | 93.22 | 93.22 |
| Naïve Bayes | 73.92 | 73.92 | 75.84 | 73.59 |
| Random Forest | 85.83 | 85.83 | 86.24 | 85.53 |

Table 6: Results of topic prediction using lyrics and model extracted theme topics

# Bibliography

* Ajoodha, R., Klein, R. & Rosman, B., 2015. *Single-labelled music genre classification using content-based features.* s.l., IEEE.
* Bahuleyan, H., 2018. *Music Genre Classification using Machine Learning Techniques.* s.l.:arXiv.
* Bužić, D. & Dobša, J., 2018. *Lyrics classification using naive bayes.* s.l., s.n., p. 1011–1015.
* Prajwal, R., Sharma, S., Naik, P. & Mk, S., 2021. Music genre classification using machine learning. *Int. Res. J. Mod. Eng. Technol. Sci,* Volume 3, p. 953–957.