Project Report

on

Telugu lyrics based classification by Using Naive Bayes

Submitted by

Sk.Mohaseen

Under the guidance of

M Muni Babu

M.Tech, (Ph.D.), Assistant Professor

Department of Computer Science and Engineering

Rajiv Gandhi University of Knowledge and Technologies(RGUKT),R.K.Valley, Kadapa, Andra Pradesh.

Rajiv Gandhi University of Knowledge Technologies
RK Valley, Kadapa (Dist), Andhra Pradesh, 516330

\\

CERTIFICATE

This is to certify that the project work titled **"Telugu lyrics based classificaation by Using Naive Bayes"** i*s a bonafied project work submitted by* ***Sk.Mohaseen*** *in the department of COMPUTER SCIENCE AND ENGINEERING in partial fulfillment of requirements for the award of degree of Bachelor of Technology in Computer science and engineering for the year 2021-2022 carried out the work under the supervision*

GUIDE	HEAD OF THE DEPARTMENT
M MUNIBABU	*P.HARINADHA*

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts success.

I am extremely grateful to our respected Director,Prof. K. SANDHYA RANI for fostering an excellent academic climate in our institution.

I also express my sincere gratitude to our respected Head of the Department Mr.P.HARINADHA for his encouragement, overall guidance in viewing this project a good asset and effort in bringing out this project.

I would like to convey thanks to our guide at college Mr. M. MUNI BABU for his guidance, encouragement, co-operation and kindness during the entire duration of the course and academics.

My sincere thanks to all the members who helped me directly and indirectly in the completion of project work. I express my profound gratitude to all our friends and family members for their encouragement.

INDEX

S.NO	***INDEX***	***PAGE NUMBER***
1	*Abstract*	*1*
2	*Introduction*	*2*
3	*METHODS AND MEASURES*	*3-5*
4	*EXPERIMENT AND RESULTS*	*5-11*
5	Conclusion	*12*
6	References	*12*

Abstract

The main objective of this project is to predict the song category by using a Naive Bayes classifier. This project describes how Naive Bayes holds the classification results more accurate with the whole song given as input to the model. This mainly holds the concepts of data pre- processing, feature extraction, and text classification to evaluate the model. The dataset consists of lyrics collected from four different genres, such as Melody, Sad, Rainy, and Pelli (marriage).

This proposed method performs classification and calculates accuracies for the given dataset. The final accuracy obtained for this model is 92.3% using the Naive Bayes classifier.

I.__Introduction__

Text classification is one of the vital tasks in supervised machine learning , which is used to assign some tags or categories based on the given texts or documents. Text classification is the heart of all variety of software which process text data. One of the primary functions of text classification is to assign the categories automatically to the songs based on lyrics by using Natural Language Processing . Text classification can be used to categorize web pages, library e-books, gallery images, news articles so on. The main application is in email spam detection , a sentimental analysis. Text classification uses many classifiers for a successful classification.

Nowadays, the heap of music on the internet is gradually increasing day by day, and organizing this vast data is an arduous task. So for a given massive data, the classification should be made automatically. For this, we need some essential concepts in data mining like Text classification, Naive Bayes. Audio features such as tempo, rhythm, pitch or lyrics features such as word length, word frequencies, sentence, and phrase structure, etc. are used to classify songs. In today's World, there exist many websites, software, apps that are available as a source to a huge set of songs. Recently, a method of recommending songs according to a particular emotion, movies, music genres have become famous. This type of work can be done quickly by using concepts like text classification.

It mainly deals with the problem to find out whether the classifier automatically predicts the categories of the songs based on its lyrics. We have chosen a set of songs from the respective category and then trained the model the results shown were based on the more frequent words in the respective category.

Accuracy in category prediction when Support Vector Machine's was applied varied from 67% to 97%. When KNN was used accuracy varied from 50% to 65% depending on the category and when the EPOCH concept used varied from 60% to 97%. This research paper is organized as follows. and Section 2briefly describes the methods and measures, Section 3 describes experiments and results.

II. METHODS AND MEASURES

The Naive Bayes classifier is a simple classification technique that assumes that each attribute is independent of each other. Naive Bayes is a machine learning algorithm whose classification efficiency is proved in applications such as document categorization and e-mail spam filtering. This classifier learns through a document classification algorithm, moreover, is based on simple usage of the Bayes rule:

Where in: $P\left(\frac{C}{T}\right)=\frac{P\left(\frac{T}{C}\right)P(C)}{P(C)}$

- C is a class,
- T is a text,

P(C) is a class probability,

- P(T) is a prior probability of text,
- P(T/C) is the conditional probability of the text for the given class,
- P(C/T) is the conditional probability of text T belongs to the class C.

Naive Bayes classification has the great advantage of computational efficiency because it has fewer computations required when compared to the remaining classification algorithms in modeling and predicting the classes, especially for large datasets. Naive Bayes can resist to over fitting and can handle a large number of attributes without the need for their selection.

One of the significant limitations of Naive Bayes is the assumption of independent factors. In real life, it is almost impossible that we get a set of entirely separate elements.

A. Performance measures

After creating a machine learning model, it is necessary to measure model performance to decide if the model is satisfactory, or whether it can be improved or even discarded.

To know the performance, we have the performance measures like confusion matrix, accuracy, and precision.

Confusion matrix [Table I]: It is used to visualize the performance in the matrix form; that is how many of the classes are classified correctly and how many of the classes are miss-classified. It has four metrics, i.e., True Positive (TP), True Negative (TP), False Positive (FP) and False Negative (FN). A True Positive is a predicted outcome where the model correctly predicts the actual class. Similarly, a True

TABLE I
CONFUSION MATRIX:

		actual	
		yes	***no***
predicted	YES	TP	FP
	No	FN	TN

Negative is a predicted outcome where the model accurately predicts the negative category. A False Positive is a predicted outcome where the model incorrectly predicts the actual class. A False Negative is a predicted outcome where the model incorrectly predicts the negative category. In the confusion matrix, the rows represent the classifier decisions, and the columns represent the actual decision.

Precision is defined as the proportion between the True positive((1))
outcomes and total predicted positive outcomes and formula referred to as below:

$$Precision = TP/(TP + FP) \qquad (2)$$

The recall is defined as the proportion of True Positive outcomes and total actual positive outcomes and formula

Referred as below:

$$Recall = TP/\ (TP + FN) \qquad (3)$$

Accuracy is defined as the proportion of total correctly classified outcomes over all outcomes and formula referred as below:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \qquad (4)$$

Error is defined as the proportion of total incorrectly classified outcomes over all outcomes and formula referred to as below:

$$Error = (FP + FN)/(TP + TN + FP + FN)\ (5)$$

Or simpler:

$$Error = 1 - precision \qquad (6)$$

F-1 measure is a combination of precision and recall. It represents the harmonic mean of the precision and recall and formula referred to as below:

$$F - 1 = ((2 precision recall)/(precision + recall) \quad (7)$$

The F-1 measure is a single-number measure used in natural language processing, information retrieval, and machine learning [14].

III.EXPERIMENT AND RESULTS

In this research paper, all the tests were performed under the environment of Intel® Core TM i5-3230M CPU @ 2.60GHz
× 4 with 4096 MB RAM, and Ubuntu 16.04 LTS 64 bit operating system. The model was implemented in python by using Machine Learning package called scikit learn .

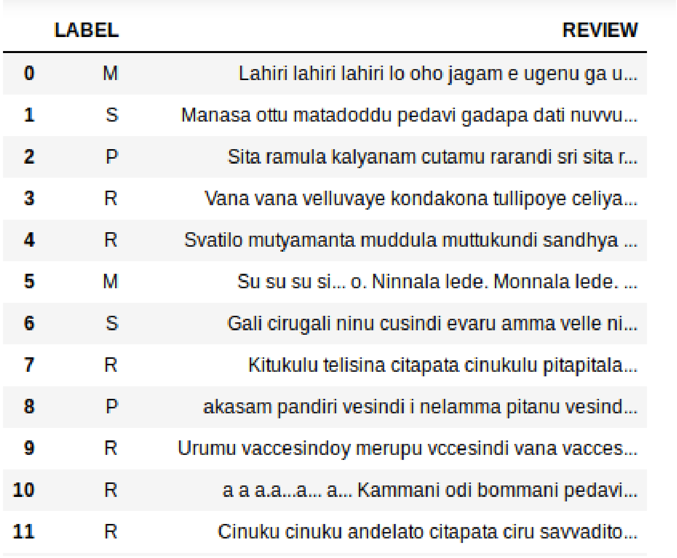
This research aims to find out whether the selected classifier can correctly identify the type of the given song based on lyrics. A set of data collected for research, the data was pre-processed and structured to give as input to the model. Consequently, the model was trained, evaluated, and the last step was a model improvement. An end-to-end pipeline text classification is composed of three main components:

A. Data preparation and preprocessing

The research question in this project was whether a classifier can classify the song category based on only lyrics.

The first step in the Data Preparation is to create the dataset. All songs were collected from the website. We can download the lyrics of a particular song from this website. Later the lyrics of Telugu songs were converted into English script through Google Translator to get the same kind of spellings, which means we are not taking the exact English meaning to the words that are appeared in the Telugu lyrics however, we are trying to get the spelling of Telugu word

fig2:



	LABEL	REVIEW
0	M	Lahiri lahiri lahiri lo oho jagam e ugenu ga u...
1	S	Manasa ottu matadoddu pedavi gadapa dati nuvvu...
2	P	Sita ramula kalyanam cutamu rarandi sri sita r...
3	R	Vana vana velluvaye kondakona tullipoye celiya...
4	R	Svatilo mutyamanta muddula muttukundi sandhya ...
5	M	Su su su si... o. Ninnala lede. Monnala lede. ...
6	S	Gali cirugali ninu cusindi evaru amma velle ni...
7	R	Kitukulu telisina citapata cinukulu pitapitala...
8	P	akasam pandiri vesindi i nelamma pitanu vesind...
9	R	Urumu vaccesindoy merupu vccesindi vana vacces...
10	R	a a a.a...a... a... Kammani odi bommani pedavi...
11	R	Cinuku cinuku andelato citapata ciru savvadito...

in English. Google Translator will do this work quickly . The result is shown in Fig. 2.

Text pre-processing for this particular dataset includes HTML decoding, remove stop words and white spaces, changing the text to lower case, remove punctuation, remove bad characters, and finally stemming.

After pre-processing, the dataset can be used further. See the sample of the dataset before and after transformations in Fig. 3.

The dataset consists of two columns (song type and song lyrics) and 120 rows (30 Melody's, 30 Sad's, 30 Rainey's, 30 Pelli's (marriage) songs). The first column, 'type' contains one-letter information about to which category it belongs to ('M' for Melody,' S' for Sad,' R' for Rainy, or' P' for Pelli (marriage)). The second column contains song lyrics.

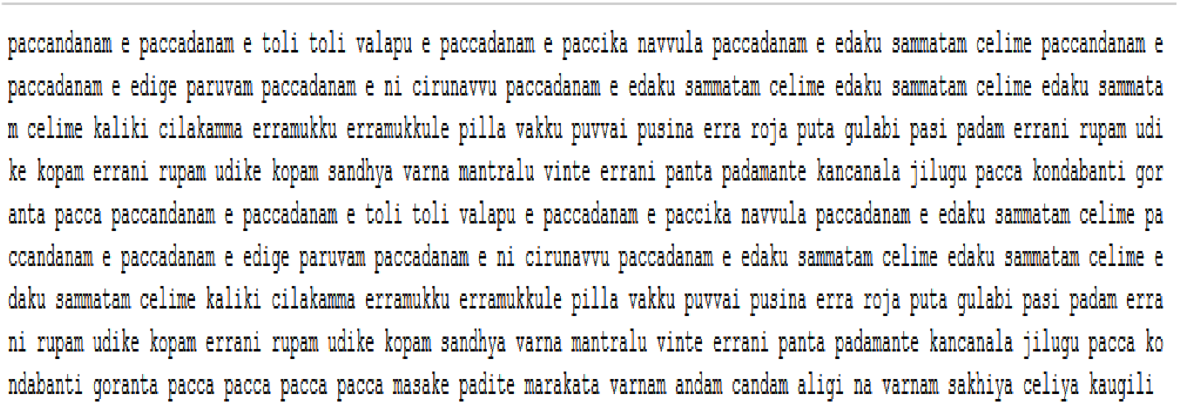
After the initial dataset creation, the randomization of rows was made. The dataset is then split into train and testing data. Since at the end of dataset creation, there was a known share of songs (90% of songs for the training set(learning) and 10% of the songs for the test set). In the end, once again, the training set and test set were separately randomized. Two-third of data

(108) would be used for learning and the remaining one-third for testing. .

B. Feature Engineering

In this Feature Engineering step, we are extracting the best feature that makes our classification successful. In this, the raw dataset is transformed into flat features that can be applied in a machine learning model.

This step includes the process of creating new features from the existing data. In this project, we are considering the frequencies of the words from the respective category and the word length, i.e., here we can take the words which are occurring very frequently and the words that have more than



```
paccandanam e paccadanam e toli toli valapu e paccadanam e paccika navvula paccadanam e edaku sammatam celime paccandanam e
paccadanam e edige paruvam paccadanam e ni cirunavvu paccadanam e edaku sammatam celime edaku sammatam celime edaku sammata
m celime kaliki cilakamma erramukku erramukkule pilla vakku puvvai pusina erra roja puta gulabi pasi padam errani rupam udi
ke kopam errani rupam udike kopam sandhya varna mantralu vinte errani panta padamante kancanala jilugu pacca kondabanti gor
anta pacca paccandanam e paccadanam e toli toli valapu e paccadanam e paccika navvula paccadanam e edaku sammatam celime pa
ccandanam e paccadanam e edige paruvam paccadanam e ni cirunavvu paccadanam e edaku sammatam celime edaku sammatam celime e
daku sammatam celime kaliki cilakamma erramukku erramukkule pilla vakku puvvai pusina erra roja puta gulabi pasi padam erra
ni rupam udike kopam errani rupam udike kopam sandhya varna mantralu vinte errani panta padamante kancanala jilugu pacca ko
ndabanti goranta pacca pacca pacca pacca masake padite marakata varnam andam candam aligi na varnam sakhiya celiya kaugili
```

Fig . . Data after Text Mining

TABLE II

FREQUENCY TABLE OF WORDS IN A DATA

Melody	***freq***	***Rainy***	***freq***	***Sad***	***freq***
kala	11	citapata	11	alayana	4
vila	4	tadise	15	Prema	17
ravali	4	jallula	7	Ani	21
cirugali	5	vana	49	Ravani	4
alo	5	cinuku	12	Amma	16
prema	10	tullipoye	4	Maunan	5
navvu	8	edo	17	nammaka	4
navve	10	svati	11	Ila	13
unna	9	meghapu	7	unnadi	7
nuvvu	13	urumu	10	Aina	19
paccada nam	12	kammani	7	Mata	12
lahiri	15	hayi	15	kadu	8
bagunda	18	cali	11	Ninu	9
ananda	12	dosili	4	tappani	4
subbala ksmi	33	ratiri	4	prati	11

navve	11	bommani	4	unna	11
roju	10	hoy	7	madi	11
tandane	24	mutyapu	7	Nuvvu	17
bagundi	12	adigenu	4	velutunna	22

a particular length in a song respectively. For example, see Table II, which is having “navvu” as the frequency of “8” in the Melody category and each word is more than a length of 3, as you see above in table II.

By observing Table II, the threshold of frequent words was set to 3. The words which have length more than 3 have taken into consideration. The words which are less than the threshold were removed. The model has shown the best results when the threshold is set to 3 or 4. By changing the threshold value, the classifier made more incorrect decisions.

In the next step of the implementation, we are trying to train the model by the existing dataset.

Melody and Sad respectively. Test dataset classified with an accuracy of 69.2%.

C. Model Training

The final step is the model building step in which a data mining model is trained on a labeled dataset. In this step, we are using a Naive Bayes classifier to train the model. Randomization makes the model learn more about train data. So here we are using “epoch” conditions to train the model correctly. Epoch is a concept in which we are continuously randomizing the data so that the model can learn more about the train data. The model training step is essential to get more accurate results and for successful classification.

D. Results and evaluation

The results of classification by using the Naive Bayes classifier without Epoch for the training dataset are shown in a confusion matrix in Table III. The training dataset classified with an accuracy of 100%.

TABLE IV:
CONFUSION MATRIX OF TEST DATASET:

predicted	M	1	0	0	3
	P	1	4	0	0
	R	0	0	2	0
	S	0	0	0	2

Maybe those miss-classifications (Table IV) happen due to the words which exist in the Melody songs are more familiar with the words in sad songs, that’s why model classified incorrectly. For the evaluation measures computing, we can use confusion matrix results, and the result is shown below :

The Accuracy is (1+4+2+2)/(1+4+2+2+3+1)=9/13= 0.692307692

The error is 1-0.692307692 = 0.307692308

The precision for class melody is 1 / (1 + 1) = 1/2 = 0.5 the recall for class melody is= 1 / (1 + 3) = 1/4 = 0.25

The F-1 measure for class melody is F1 = ((2 × 0.5 × 0.25) / (0.5 + 0.25) = 0.25/0.75 = 0.33333

E. Model improvement

The results of the classification by using the Naive Bayes classifier along with the epoch concept for the training dataset are shown in a confusion matrix in Table V. The training dataset is correctly classified with an accuracy of 100% when the epoch value is 15. It can be noticed that accuracy for the training dataset with and without epoch is the same.

The Epoch concept was applied to improve the model efficiency. The model gave the best results for the test dataset

TABLE V

CONFUSION MATRIX OF TRAINING DATASET FOR IMPROVED MODEL

		Actual			
		M	***P***	***R***	***S***
predicted	M	28	0	0	0
	P	0	28	0	0
	R	0	0	26	0
	S	0	0	0	27

TABLE VI

CONFUSION MATRIX OF TEST DATASET FOR IMPROVED MODEL

		actual			
		M	***P***	***R***	***S***
predicted	M	2	0	0	0
	P	0	4	0	0
	R	0	0	3	1
	S	0	0	0	3

When Epoch value is 15. With this adjustment, the classifier recognized all songs correctly, but only one song was miss- classified (Table VI).

After the model improvement, the results are as follows (From table VI):

The accuracy is (2+4+3+3) / (2+4+3+3+1)=12/13= 0.9230769230769231

The error is 1 – 0.9230769230769231 = 0.076923077

The precision for class melody is 2 / (2 + 0) = 2/2 = 1.0 the recall for class melody is= 2 / (2 + 0) = 2/2 = 1.0

The F-1 measure for class melody is F 1 = ((2 × 1.0 × 1.0) / (1.0 + 1.0) = 2/2 = 1.0

So, finally we can achieve up to 92.3% of accuracy with an improved model.

Finally, we can see precision and recall values for each class precisionand recall values somewhat less than M and p so that we can consider the respective precentage as a result.

In table VII, precision, recall, F-1 score for Melody (M), Pelli (marriage) (P), are 1.00, which means those two classes are correctly classified. Rainy (R) and Sad (S) classes have

	Precision	***Recall***	***F-1 score***
M	1.00	1.00	1.00
P	1.00	1.00	1.00
R	1.00	0.75	0.86
S	0.75	1.00	0.86

IV.Conclusion

Creating a dataset was a tedious and time-consuming task, partly because it was created manually and partly because of doubt about inserting some songs into the dataset. Namely, cases such as some songs belong to more than one category. In those cases, we can categorize that song with multi-labels as output. The results of a created model were perfect. Naive Bayes classifier is the right choice for this task – once again Naive Bayes proved its capabilities. Since the dataset was quite small, it was a logical candidate for the model. The result showed that Melody, Sad, Pelli (marriage), and Rainy songs have textual 'signatures' that can be distinguished in no small degree solely on reading the text.

References

[1]G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[2]V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes?," 2006.

[3]Sadovsky Adam and Chen Xing, "Song genre and artist classification via supervised learning from lyrics," *Previous CS224N Final Project*, 2006.

[4] K. Choi, J. H. Lee, and J. S. Downie, "What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 453–454, IEEE Press, 2014.

[5]D. Bu žic´ and J. Dobsˇa, "Lyrics classification using naive bayes," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1011–1015, May 2018.

[6]Y. Xia, L. Wang, and K.-F. Wong, "Sentiment vector space model for lyric-based song sentiment classification," *International Journal of Computer Processing Of Languages*, vol. 21, no. 04, pp. 309–330, 2008.

[7]H. Abburi, E. S. A. Akkireddy, S. Gangashetti, and R. Mamidi, "Multi- modal sentiment analysis of telugu songs.," in *SAAIP@ IJCAI*, pp. 48– 52, 2016.

[8]J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*.
Elsevier, 2011.

[9]F. Caraffini, "The naive bayes learning algorithm," 02 2019. [13]Y. Yang and T. Joachims, "Text categorization."

[10]D. M. Powers, "What the f-measure doesn't measure: Features, flaws, fallacies and fixes," *arXiv preprint arXiv:1503.06410*, 2015.

[11]"https://scikit-learn.org/stable/." [16]"www.telugulyrics.org." [17]"https://translate.google.co.in."

[12]"Confusion matrix, https://www.geeksforgeeks.org/confusion-matrix- machine-learning/."

[13]J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[14]S. R. S. Dokkara, S. V. Penumathsa, and S. G. Sripada, "Morphological generator for telugu nouns and pronouns," *International Journal of Computer Applications*, vol. 165, no. 5.