



Marathi Poem Classification using Machine Learning

R. A. Deshmukh, Suraj Kore, Namrata Chavan, Sayali Gole, Kumar Adarsh

Abstract: Poem a piece of writing in which the expression of feelings and ideas is given intensity by particular attention to diction (sometimes involving rhyme), rhythm, and imagery. It is used for showing different views. Every poet writes a poem with a different intention and different views. In the proposed system we have classified the poem according to its sentiments by using words of different categories. Machine learning algorithm SVM classifier is used for differencing the class of the poem. This system also enables the user to search the poem based on the poet name and poet type. For 341 poems of five categories 'Friend', 'Prem', 'Bhakti', 'Prerna' and 'Desh' accuracy achieved is 93.54%.

Index Terms: Machine Learning, Support Vector Machine, sentiments, Confusion matrix, Accuracy, Error rate.

I. INTRODUCTION

Poetry is kind of expressive combination of lines and words that makes uses of beauty and rhythmic quality of language. The poems have many sentiments to derive from its line and are often very specific towards something. Poems usually have different meaning and it is usually written to depict some object or thing.

It has rich tradition and mostly used language for expressing something. In the proposed system we are classifying the poem according to its sentiment based on words category and we are working with Marathi language we have seen the work done by poets like Mangesh Padonkar, Arun Kolatkar etc. in poetic gem tradition. In the poetry analysis basic we generally focus on line and words for classification purpose.

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

Prof. Rushali A. Deshmukh*, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, affiliated to Savitribai Phule Pune University, Pune, India.

Mr. Suraj Kore, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, affiliated to Savitribai Phule Pune University, Pune, India.

Ms. Namrata Chavan, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, affiliated to Savitribai Phule Pune University, Pune, India.

Ms. Sayali Gole, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, affiliated to Savitribai Phule Pune University, Pune, India.

Mr. Kumar Adarsh, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, affiliated to Savitribai Phule Pune University, Pune, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license [http://creativecommons.org/licenses/by-nc-nd/4.0/](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Each line of the poem contains all the mood related terminologies in the poems that are required to identify based on the classification of the poem.

Poem classification is done by using the SVM algorithms and the terminology. From the past, the poems are very popular in the world. Every poem has quality meaning in it. We have developed the project for the identification of the poem and classification into type of the poem; we are motivated to this project on this basis. By classifying the poem and we are able to understand the type of poem and poet general nature. In this approach, we classify the poem according to its sentiment. This project depicts poet identification, poem classification (love, nature, etc.), Search by poet name and by poem sentiments. To build a system i.e. "Marathi Poem Classification Using Machine Learning".

II. LITERATURE REVIEW

"Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffari" used subject based classification using its syntactic and semantic features of Bangla poem and used stylometric features to find poems orthographic features, syntactic features, etc. [1]. We are going to refer to which factors they have used for classification the poem and finding its orthographic, syntactic and lexical feature. "Navinder Kaur, Amandeep Verma" performed authorship attribution to find author from unseen text and used a classification algorithm to find the author of testing data and analyzed accuracy, precision, recall factors[2]. From this paper, we have learned how to train the classifier model and how to evaluate its performance.

"Pandian*, V. V. Ramalingam and R. P. Vishnu Preet" used author work in the past for classification of unknown Tamil poem and used a decision tree algorithm to find classifier accuracy [4], and we are going to implement it in a better way for our system. "Zhou Guo Dong" used the MIIM model and maximum entropy and chunking strategy to find unknown words [7], we are going to prefer the strategy used in it for checking its performance.

For unknown word detection "Xia Li, Bin Wu, Bailing Zhang" have used the word embedding technique for classification and use of active distance [3]. The results are stored in the dictionary to analyze its meaning. "Jasleen Kaur, Jatinder Kumar R. Saini" uses the concept of NLP, Machine learning, KNN for classification of Punjabi poem [5]. Results show that each classification algorithm has different accuracy; we will analyze the same for our system and use the best algorithms. The work in Indian regional languages like Tamil, Punjabi, and Bangla provided us a brief idea about poet identification, poem summarization, and unknown word detection through various techniques and classification algorithms. We have referred various methods for author and poem detection like we can detect an author using the known work of the author [8].

The chunking strategy and feature based extraction used in gave some briefing about unknown word detection and its analyzation according to the poet point of view [6][7]. The emotion classification methods used can be used in sentiment analysis according to poet nature and their view as it is not done in many of the regional languages of India [9]. Further Automatic metric detection can be done using the SVM classifier [10], and we will try to remove the incorrect detection in the future implementation of our idea.

III. PROPOSED SYSTEM

Fig. 1 shows the proposed system for Marathi Poem Classification.

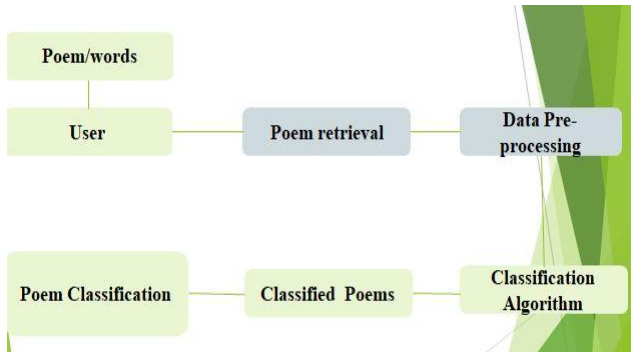


Figure 1: Proposed system

In the proposed system we have taken words in our dataset then we have separated stop words and trained the words using Support Vector Machine algorithm, then by using those trained words we classify the poem and to our dataset. The classified poems are according to their sentiments. Using the poem in data we can have search by poet and search by poet sentiment facility.

Fig. 2 displays how the system will work.

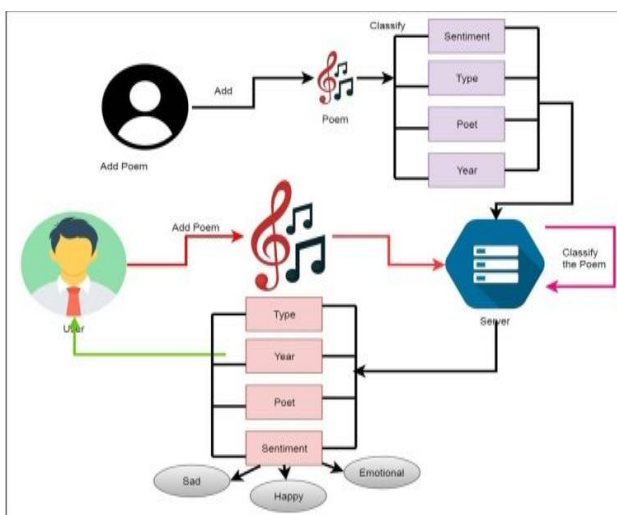


Figure 2: System Architecture

The proposed system will enable the user to find a poem based on category wise classification and poet based classification. The user will understand the different nature of the poem and unique feature, with the help of words which categorize poem into different category like nationalism, love, etc. Users will understand the peculiar feature of every category. The proposed system will be

beneficial to the user who has little understanding of Marathi language. It will also enhance poet reputation by analyzing their work in Marathi language and their style of writing the poem and enhancing their style of poetry.

IV. ALGORITHM

“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

DATABASE

- The initial dataset contains at least 200 words for each category like love, friendship, nationality. These words are used to train the SVM classifier. This word dataset can be increased in the future with the increase in poems.
- We are storing this data in a structured format using MySQL.
- We have a stopwords dataset (the words which are having no sentiments).

Algorithm: Poem Classification

Input: poem

Output: category of poem

Initialize: an array array[] with words of poems

1. array[]=split(poem);
2. for(word:array)
3. if(Ispresent(word,stopword))
4. Tfidf<<category>>remove(word,array)
5. Array presentword[category]=0;
6. Map<<string,int>> tfidf;
7. for(word:array)
8. If(category=Ispresent(word,prime))
9. Tfidf[category]++;
10. max=0;
11. for(i=0;i<category.length;i++)
12. if(tfidf[category]>max)
13. return max;

- Input: Poem
- Poem words are split with respect to space and taken in an array.
- We are referring stop words dataset to remove stopwords from the array.
- Now we are referring word data set to recognize the category of words that are present in the array.
- The category wise word data is added to the demo table.
- The SVM classification algorithm is used.
- The tfidf(frequency of a word is calculated) of each category is calculated.
- Maximizing the distances between the nearest data point of each class will help us to decide the right hyper-plane and classes are separated.

- Now when we add an untrained poem then a poem is classified according to its word sentiments.

There are three modules in our system.

1) Module 1: Adding words and stop words to the dataset

In this module, we are adding stop words to stop word dataset, which can be removed at the time of poem extraction.

Then we are adding sentiment wise words to the word dataset, which can be further used to train the classifier model.

2) Module 2: Poem Classification

- Poem words are split with respect to space and taken in an array.
- We are referring stop words dataset to remove stop words from the array.
- Now we are referring word data set to recognize the category of words that are present in the array.
- The category wise word data is added to the demo table.
- The SVM classification algorithm is used here.
- The tfidf (frequency of a word is calculated) of each category is calculated.
- Maximizing the distances between the nearest data point of each class will help us to decide the right hyper-plane and classes are separated.
- Now when we add an untrained poem then the poem is classified according to its word sentiments.

3) Module 3: Search Module

a) Search by poet name.

In this module, we are firing a query to search poems related to a particular poet.

b) Search by poem sentiment.

In this module, we are firing a query to search poems related to the particular poem category.

V. MATHEMATICAL MODEL

Let us consider S be a Systems such that

$$S = \{Ad, U, R, BN, P, \}$$

U=User

P=Prediction Result

R=Recommendations

S=System

$$U = \{U_1, U_2, U_3, \dots, U_n\}$$

There may be a number of users for making use of the system. So this is the Infinite Set.

$$R = \{R_1, R_2, R_3, \dots, R_n\}$$

It is the technique applied to the given parameters. So this is a finite set.

• EVENT 1

User will make registration on System & Storage Server.

Let f(U) be a function of User

Thus, f(U) -> {Ss}

• EVENT 2

User will be authenticated.

Let f(A) be a function of System

Thus, f(A) -> {U1, U2, U3, ..., Un} e Ss

VI. EXPERIMENTAL SETUP

The initial dataset contains atleast 50 words per category like love, friendship, nationality. These words are then trained using the SVM algorithm. We have added stop words as well in our data. Poems are taken as input, it is added into an array by splitting the poem. The special characters are removed. The poems which are classified into the particular category are added into the dataset and it is used further for finding accuracy.

Table 1: Category wise word count

Category	Number of words
Friend	230
Prem	250
Bhakti	200
Prerna	250
Desh	240
Total	1170

Category wise number of words used for classification are as shown in Table 1. From that most useful words are shown in Table 2.

Table 2: Category wise most useful words

Category	Most Useful Words
मैत्री (Friend)	मैत्री, मित्रासाठी, मित्रामुळे, मित्रांकरिता, मित्राबरोबर, संगती, मैत्रीण, मैत्रिण, मैत्रिणीसाठी, मैत्रिणीमुळे, फ्रेंड, फ्रेंड्स, मैत्रीचं, मैत्रीविना, मैत्रीस, मैत्रीचा, मैत्रीसाठी, मैत्रीच्या, मित्र, मित्रांवर, मैत्रीत, मित्रांची' etc.
प्रेम (Prem)	प्रेम, प्रेमासाठी, प्रेमकरता, प्रेमाकडे, प्रेमामध्ये, प्रियासी, प्रियकर, प्रेमगंध, प्रेमिका, प्रेमात, प्रियकर, प्रियतम, प्रियतमा, मन, मनात, मनाला, मनासाठी, मनामध्ये, हृदय, प्रेमाचा, प्रेमाचे, हृदयाचा, हृदय, विश्वास, साथ' etc.
भक्ती (Bhakti)	गणपती, मंगलमूर्ती, किर्ती, विठ्ठला, पांडुरंग, अभंग, तिर्थ, तुकाई, विठाई, माऊली' etc.
प्रेरणा (Prerna)	दुनियादारी, सहन, धडपडत, मेहनतीच्या, अडथळे, ध्येयाच, झेलण्यासाठी, लढाई, उड्डाण, जिद्द, मनात, परीक्षा, ताकदीची, सळसळते, रक्त, हुरलीय, ध्येयवेडा, हिम्मत, अबला, उमेद etc.

देश (Desh)	देशभक्तीचा, स्वातंत्र्य, राष्ट्र, प्रमाणिकतेला, भ्रष्टाचारा, स्वातंत्र्याची, बलीदान, इतिहासाचे, न्याय, हक्क, हिंदुस्तान, विजय, क्रांती, शांती, एकात्मतेची, वीरांचा, इतिहासा, सद्वादीच्या, स्वराज्य, आजाद, बंधुभाव, राष्ट्रभावना, भारतीयत्वाच, बलसागर, सीमोल्लंघन etc.
---------------	---

For the implementation part we have used the SVM classification algorithm which is a supervised learning algorithm of Machine Learning. It is used for classification of the poem according to its sentiments and help the user to understand the poem.

Generalize steps in the project are as follows:

- Loading of words of different feature
- Poem Classification
- Search by Poet.
- Search by poem sentiments
- Share Result and Data

Category wise poems used for our experimentation are shown in Table 3.

Table 3: Category wise Poem count

Category	Number of Poems
मैत्री (Friend)	78
प्रेम (Prem)	82
भक्ती (Bhakti)	65
प्रेरणा (Perna)	65
देश (Desh)	51
Total	341

VII. RESULTS

- Poem Classification is done according to its sentiments.
- User and admin both have the functionality of adding the poem, search by poet, and search by poem sentiments.
- Admin has additional functionality of accepting and rejecting the poem that the user has added.
- The poem is classified according to the word trained in the dataset.
- The words in the poem are labeled according to its respective feature and feature whose frequency count is more the poem is classified in that feature.
- Till now we have added 341 poems in the dataset.
- Results or accuracy depend on training of high number of words in the dataset.

Table 4 shows the Confusion Matrix for poem classification in which column indicates predicted class and row indicates actual class. For e.g. for the 'Prem' category using our classifier 80 poems are classified in that category, out of that 75 poems are correctly classified in the 'Prem' category. Diagonal values represent correctly classified poems count.

Table 4: Confusion Matrix

Category	देश	प्रेम	मैत्री	प्रेरणा	भक्ती	Total
देश	45	2	0	3	1	51
प्रेम	0	75	4	2	1	82
मैत्री	0	3	73	0	2	78
प्रेरणा	0	0	0	64	1	65
भक्ती	0	0	0	3	62	65
Total	45	80	77	72	67	341

Category wise Precision, Recall, F1-score, and Support is as shown in Table 5.

Table 5: Category wise precision, recall, f1-score and support

Category	Precision	Recall	F1-score	Support
देश	1.00	0.88	0.94	51
प्रेम	0.94	0.91	0.93	82
मैत्री	0.95	0.94	0.94	78
प्रेरणा	0.89	0.98	0.93	65
भक्ती	0.93	0.95	0.94	65
Avg.	0.94	0.93	0.94	341

The accuracy achieved is 93.54%

VIII. CONCLUSION AND FUTURE SCOPE

In this research work, a new Poem classification and poet identification task have experimented with Marathi poetry written by several different poets. Several sentiment features have been observed and viewed for Marathi poetry. For the classification purpose, the SVM classifier is used which uses 500 words for training and 100 poems for the testing based on trained words. Experiments of poem classification and poet identification on poems of different poets, which have been done separately for each feature and sentiments, results in the following observable points:

- The word-based features are easy and simple to classify.
- Failure occurs when the words present in the poem are not trained with their label.

- The features are evaluated on the basis of the words trained in the database.
- Although the size of the words was too small the performance of the classifier is really interesting.

This work of Poem Classification, which is one of the rare works done on the Marathi language, shows a real motivation and interest for this language. System when tested on 341 poems accuracy achieved is 93.54%.

We look forward using the bigger dataset to improve accuracy, considering the internationalization.

As we are implementing our system for the poem, we can increase the scope of the system to the story, essay.

We can use Deep learning in our idea for future implementations.

The challenges in our model are:

- a) The same word can be treated as Desh, bhakti or Prerna depending on context thus making it difficult to identify the real context of text.
- b) Meaning of words change with respect to poet thinking.
- c) Difficult to identify ambiguous poem.
- d) Marathi Language should be known by user.
- e) Internationalization

REFERENCES

1. Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffari "Automated Analysis of Bangla Poetry for Classification and Poet Identification" Journal IICA, 2015
2. Navinder Kaur, Amandeep Verma "Authorship Attribution of Punjabi Poetry using SVM Classifier" Journal IJARCSE, May 2015
3. Xia Li, Bin Wu, Bailing Zhang "Unknown Word Detection in Song Poetry". 2016 IEEE First International Conference on Data Science in Cyberspace.
4. Pandian*, V. V. Ramalingam and R. P. Vishnu Preet" Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using Bayes Net Algorithm". Indian Journal of Science and Technology, December 2016
5. Jasleen Kaur, Jatinder Kumar R. Saini "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms". Journal ICMIC, 2017
6. Zhongshilte, Wen-Ting Liang "SVM based classification metrics for poetry style". International Conference on Machine Learning, 2007.
7. Zhou Guo Dong "A Chunking Strategy towards Unknown Word Detection in Chinese Word Segmentation". Institute for Infocomm Research, 2016.
8. Dr. Binoy Barman! "A contrastive analysis of English and Bangla phonemics". The Dhaka University Journal of Linguistics: Vol. 2 No.4 August 2009 Page: 19-42, Published on August 2010 by the Registrar, Dhaka University ISSN N-2075-3098
9. Ouais Aishray, deema Alshamaa, Nada Ghrem "Emotion Classification in Arabic Poetry using Machine Learning". International Journal of Computer Applications (0975 – 8887) Volume 65– No.16, March 2013
10. Saeid Hamidi, Farbod Razzazi" A Meter Classification System for Spoken Persian Poetries". Signal Processing-An International Journal (SPIJ), Volume (5) : Issue (4) : 2015.

AUTHORS PROFILE



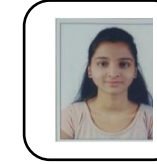
Mrs. Rushali Deshmukh (Dhumal) pursued Bachelor of Computer Engineering from Pune University, India in 1999 and Master of Computer Engineering from Pune University, India in the year 2007. She is having 19 years of teaching experience. She has published/presented 24 papers in reputed Journals/Conferences. She has registered copyright on "Marathi Sentiment Dataset" and also published a patent "Improved smart shopping cart using client-server". She is currently pursuing a Ph.D. in the area of Semantic Analysis of Natural Languages. Her main research work focuses on Natural Language Processing using Machine learning and Data Mining.



Mr. Suraj Kore pursued a Bachelor of Computer Engineering from Savitribai Phule Pune University, India in 2019. He has published the paper "Survey of Marathi Poem Classification Using Machine Learning" in UGC approved Journal.



Ms. Namrata Chavan pursued a Bachelor of Computer Engineering from Savitribai Phule Pune University, India in 2019. She has published the paper "Survey of Marathi Poem Classification Using Machine Learning" in UGC approved Journal.



Ms. Sayali Gole pursued a Bachelor of Computer Engineering from Savitribai Phule Pune University, India in 2019. She has published the paper "Survey of Marathi Poem Classification Using Machine Learning" in UGC approved Journal.



Mr. Kumar Adarsh pursued Bachelor of Computer Engineering from Savitribai Phule Pune University, India in 2019. He has published the paper "Survey of Marathi Poem Classification Using Machine Learning" in UGC approved Journal. He has also published patent "Full Equipped Ruler Scale"