Automatic Marathi Text Classification

Rupali P. Patil, R. P. Bhavsar, B. V. Pawar

Abstract*: Multifold growth of internet users due to penetration of Information and Communication technology has resulted in huge soft content on the internet. Though most of it is available in English language, other languages including Indian languages are also catching up the race rapidly. Due to exponential growth in Internet users in India common man is also posting moderate size data on the web. Due to which e-content in Indian languages is growing in size. This high dimensionality of e-content is a curse for Information Retrieval. Hence automatic text classification and structuring of this e-content has become the need of the day. Automatic text classification is the process of assigning a category or categories to a new test document from one or more predefined categories according to the contents of that document. Text classification works for 14 Indian languages are reported in the literature. Marathi language is one of the officially recognized languages of Indian union. Little work has been done for Marathi text classification. This paper investigates Marathi text classification using popular Machine Learning methods such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Centroid Based and Modified KNN (MKNN) on manually extracted newspaper data from sport's domain. Our experimental results show that Naïve Bayes and Centroid Based give best performance with 99.166% Micro and Macro Average of F-score and Modified KNN gives lowest performance with 97.16% Micro Average of F-Score and 96.997% Macro Average of F-score.

The proposed work will be helpful for proper organization of Marathi text document and many applications of Marathi Information Retrieval.

***Keyword*: Automatic, Centroid Based, Classification, K-Nearest Neighbor, Marathi, Modified KNN, Naïve Bayes, Text.**

I. INTRODUCTION

With the fast growth of World Wide Web contents and users, the amount of digital information is growing exponentially. With growing multilingual e-contents language coverage has also increased. Hence there is a need for automatic text classification of web documents. In early days, most of the information available on net is in English language only.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author

Rupali P.Patil*, Department of Computer Science, S.S.V.P. S's Lk. Dr. P. R. Ghogrey Science College, Dhule, India, *rupalipatilh@gmail.com*

R. P. Bhavsar, School of Computer Sciences, North Maharashtra University, Jalgaon, India, *rpbhavsar@nmu.ac.in*

B. V. Pawar, School of Computer Sciences, North Maharashtra University, Jalgaon, India, *bvpawar@nmu.ac.in*

But in this era of digitization, information in other natural languages is also easily available on the web and increasing exponentially. People in India use different languages for their communication. Marathi is a spoken language of Maharashtra state. It is also the official language of Maharashtra and Goa states of Western India. It is one of the 22 scheduled languages of India is a spoken language of Maharashtra state. In the world, in 2001, there were 73 million Marathi speakers; Marathi ranks 19th in the list of highest spoken languages [1]. In India the fourth largest numbers of native speakers are belonging to Marathi [1]. Marathi has some of the oldest literature of all modern Indo-Aryan languages, dating from about 900 AD [2]. The major dialects of Marathi are Standard Marathi and the Varhadi dialect [3]. Malvani and Konkani have been heavily influenced by Marathi varieties.

According to government data, at the end of March, 2016, India had a total of 342.65 million Internet subscribers. Among all states in India, the highest number i.e. 29.47 million of Internet subscribers are belonging to Maharashtra state followed by the states like Tamil Nadu, Andhra Pradesh and Karnataka [4]. Thus, with the rapid growth of Internet, number of Marathi websites and number of Internet users are growing tremendously. Therefore Marathi information on web is also increased. This emphasizes the importance of applying Text mining approaches to organize this gigantic Marathi information.

Manual organization of large collection of Marathi language text documents are very difficult, time consuming and required lot of human efforts. It is advantageous to use text classification using machine learning techniques for this. It saves time, money and human efforts. Text classification is the task of classifying highly unstructured natural language text documents into one or more predefined classes based on their contents.

Different machine learning techniques such as K Nearest Neighbor (KNN) [5], Naïve Bayes (NB) [6], Decision Tree (DT) [7], Neural Network (NN) [8], N-Gram Model [9], Support Vector Machine (SVM) [10] and Centroid Based [11] are used for classification. Text classification has several applications such as Email Filtering [12] and Web Site Classification [13]. Marathi text classification can be helpful for Marathi text filtering such as Marathi Email filtering, Marathi Search Engine, Marathi web pages categorization, Marathi News Articles Organization etc. Therefore, it is motivating to design and develop automatic Marathi text classification system. This paper deals with the implementation of text classification algorithms such as KNN, MKNN, Naïve Bayes and Centroid Based for Marathi language text documents.

The rest of the paper is organized as follows: Section II summarizes Related Work; while Section III gives information about Marathi language and Issues related to Marathi text classification. Section IV describes Experimental Setup followed by Results and Discussion in section V. Section VI gives Conclusion.

II. RELATED WORK

Text Classification is an active research area of modern Information Retrieval and Machine Learning. For English and other European Languages many research papers on text classification are available. But about research in Indian Languages text classification, it is in growing stage. As Indian languages are morphologically rich and inflectional languages, classification in Indian languages text is difficult task.

Out of 22 constitutionally recognized languages and 12 Indian scripts, text classification works for 14 Indian languages are reported in the literature such as Aassamee, Bangla, Hindi, Kannada, Marathi, Malayalam, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, Thai and Urdu. According to languages this work is reported in our previous paper [14].

Recently, R. M. Rakholia and J. R. Saini [15] reported work for Gujarati Documents classification using Naïve Bayes Classifier without using feature selection and using feature selection, they reported 75.74% and 88.96% accuracy respectively. Their results proved that Naive Bayes Classifier contribute effectively for Gujarati documents classification.

Rajan, Ramalingam, Ganesan, Palanivel and Palaniappan [16] developed Tamil text classification system based on Vector Space Model (VSM) and Neural Network (NN) model. They applied inflectional rules to reduce number of terms. Their experiment using Tamil Corpus (CIIL Corpus) showed that performance of Neural Network Model (93.33%) is better than VSM (90.3%) on classification of Tamil text documents.

Narayana Swamy and Haunumanthappa [17] evaluated three text mining algorithms, K Nearest Neighbor, Naïve Bayes and Decision Tree (J48) on Kannada, Tamil and Telugu text. Their own created corpus consists of 300 documents belonging to three different categories. Stopword removal and Stemming are applied. VSM is used for document representation. Out of three Naïve Bayes (97.66%) gives best accuracy followed by Decision tree (97.33%) and then K Nearest Neighbor (93%). Their experiment proved that text mining algorithms are applicable for Indian languages text classification.

Ashis Kumar Mandal and Rikta sen [18] explored the use of machine learning approaches and analyzed the efficiency of four supervised learning methods namely Decision Tree, K-Nearest Neighbor (KNN), Naïve Bayes and Support Vector Machine (SVM) for categorization of Bangla web document. They built their own BD-Corpus with 5 different categories having 1000 documents. In pre-processing step digit, punctuation marks, stopwords are removed. Stemming is applied and TFIDF is used for feature extraction. Their experimental results show that K Nearest Neighbor and Naïve Bayes are more capable than SVM for small training set. In terms of training time SVM is fastest and decision tree (C4.5) takes more time among 4 algorithms. For sufficient training data SVM (89.44%) produces best result followed by NB (85.22%), decision tree (80.65%) and KNN (74.24%).

Mrs Sushma R. Vispute and Prof M. A Potey [19] created an intelligent System to retrieve Marathi text documents. They developed their system using Lingo Clustering Algorithm based on VSM. Their dataset consist of 3 categories having 107 Marathi text documents. Performance of Lingo Clustering algorithm is good with 91.10% accuracy for categorization of Marathi Text Documents.

Meera Patil and Pravin Game [20] presented and tested an efficient text classification system using Naïve Bayes, Centroid Based, K Nearest Neighbor and Modified K Nearest classifier. Pre-processing was performed using rule based stemmer and Marathi word dictionary. They did not use stop word list. Their experimental results of comparison of above four classifiers have been showed that Naïve Bayes is more efficient in terms of classification accuracy and classification time.

Jaydeep Jalindar Patil and Nagaraju Bogiri [21] language document retrieval system based on user profile and retrieve relevant documents which reduce human efforts. User profile contains user’s interest and user browsing history. System provides Automatic classification of Marathi text documents by using Lingo [Label Induction Grouping Algorithm] which is based on Vector Space Model [VSM]. From their experiment they found that LINGO clustering algorithm is efficient for Marathi text document categorization.

N. Dangre et.al. [22] explored depth survey of Marathi text retrieval techniques. In this paper, they proposed Marathi news retrieval system using clustering techniques and also include scope of proposed system.

Aishwarya Sahani et.al. [23] developed a Marathi text document categorization system using Lingo Clustering algorithm, as their survey reported that Lingo clustering algorithm is one of the efficient clustering algorithm for Marathi language text documents. Their automatic clustering system assigns name to the clusters based on the contents of the documents. For their experiment they have used two data sets. One general dataset having 24 general documents and another news data sets having 33 documents for news category. They reported that their system would be helpful for newspaper and news Chanel agencies to sort and organize huge information. Marathi Search Engine can also use their system.

Pooja Bolaj and Sharvari Govilkar [24] presents classification of Marathi text documents using supervised learning methods such as Modified KNN, Naïve Bayes and Support Vector Machine (SVM) and ontology based classification. Their experimental result shows that among four methods SVM gives consistently better results than other three, followed by Ontology and then MKNN and Naïve Bayes. They also compared time taken by four classifiers in terms of milliseconds and reported that as no. of testing documents increases SVM and Naïve Bayes take more time as compared to MKNN and Ontology based methods.

From the above survey it is found that not much work has been done in Marathi text classification.

III. ISSUES IN MARATHI TEXT CLASSIFICATION

Marathi is one of the officially recognized languages of India. There are total 52 alphabets in Marathi including 16 vowels and 36 consonants. Like English it is written from left to right. As compared to English, Marathi morphology is agglutinative and rich in nature.

Table-I: Various forms of Marathi word 'जगणे' (ɟəgəɳeː) (to live)

Word	Gender	Pluralities	Example Sentence
जगला *(ɟəgəlaː)(lived)*	Masculine	Singular	तो जगला *(t̪oː ɟəgəlaː) (He lived)*
जगली *(ɟəgəliː) (lived)*	Feminine	Singular	ती जगली (*t̪iː ɟəgəliː)(she lived)*
जगले *(ɟəgəleː) (lived)*	Neuter	Plural	ते जगले *(t̪eː ɟəgəleː)(They lived)*
जगले *(ɟəgəleː) (lived)*	Neuter	Singular	ते मूल जगले *(t̪eː muːlə ɟəgəleː)(That child lived)*
जगले *(ɟəgəleː) (lived)*	Neuter	Plural	ते मुलं जगले *(t̪eː muləⁿ ɟəgəleː)(Those kids lived)*
जगलीत *(ɟəgəliːt̪ə) (lived)*	Neuter	Plural	ती झाड जगलीत *(t̪iː ɟʰaːɖə ɟəgəliːt̪ə)(Those trees lived)*
जगल्या *(ɟəgəljaː) (lived)*	Feminine	Plural	त्या मुली जगल्या *(t̪jaː muliː ɟəgəljaː)(Those girls lived)*

This makes difficult to develop automatic text classification system for Marathi language text. In the context of text classification, the basic character of Marathi is similar to English, as we use the frequency distribution of content terms for both. But, the huge amount of inflections, word gender and pluralities make the pre-processing stage of Marathi text document classification system more complex than English.

The gender number person system of Marathi is same as English. Marathi also has three genders viz. masculine, feminine and neuter. Marathi words are classified into 8 parts of speeches (POS) namely Noun, Pronouns, Adjectives, Verb, Adverb, Preposition, Conjunction and Interjection, also the Marathi words have post position markers. Unlike English, Marathi words are inflected for 'case' category. Marathi has special symbols to mark the 'case' of the Marathi word using post position marker called as Vibhakti. Such feature is not present in English language.

Here we have listed some aspects of Marathi that makes automatic processing of Marathi text classification more difficult as compared to English.

Due to rich morphological nature, Marathi word is inflected for many forms. Thus handling these forms is a challenge in itself. Sample inflected forms for Marathi word जगणे_(to live) (ɟəgəɳeː) are studied in above table 'Table- I'. Marathi verb inflections are high in numbers. Sample verbs are listed as given below.

तो जगतो (*t̪oː ɟəgət̪oː)(He lives)*, ती जगते (*t̪iː ɟəgət̪eː(She lives))*, ते जगतात (*t̪eː ɟəgət̪aːt̪ə*) (They live), ते जगलेत (*t̪eː ɟəgəleːt̪ə)(They lived)*, जगतो आहे (*ɟəgət̪oː aːɦeː)(Is living)*, जगत होता (*ɟəgət̪ə ɦoːt̪aː)(Was living)*, जगत आहे (*ɟəgət̪ə aːɦeː)(Is living)*, जगणार आहे (*ɟəgəɳaːrə aːɦeː)(Is going to live)*, जगत आला आहे (*ɟəgət̪ə aːlaː aːɦeː*) (Has been lived), जगत आला होता (*ɟəgət̪ə aːlaː ɦoːt̪aː)(Had been lived)*, जगला होता (*ɟəgəlaː ɦoːt̪aː)(Had lived)*, जगत आले होते (*ɟəgət̪ə aːleː ɦoːt̪eː)(have been lived)*, जगू शकेन (*ɟəguː ɕəkeːnə*) (can live) etc.

- Due to polysemy, a word may functions different: noun, verb, preposition etc. The sample word 'जग' can appear as noun, verb and postposition. Its usages are given below

1. "हे जग विशाल आहे" (*ɦeː ɟəgə viɕaːlə aːɦeː)(This world is vast)*

   In this sentence the word 'जग' (ɟəgə) assumes the meaning as 'world' i.e. as noun POS category.

2. "पाण्याचा जग भरून आण." (*paːɳjaːcaː ɟəgə bʰəruːnə aːnə)(Bring a jug filled with water)*

   In this sentence also the same word 'जग' (ɟəgə) is used as a noun, but in second sentence the meaning of the word 'जग' (ɟəgə) is a 'pot' used to carry some water.

3. "आयुष्य नीट जग". *(aːjuʂjə niːʈə ɟəgə)(Live life well/properly)*

   In this case 'जग' (ɟəgə) is used as a verb and its meaning is to live.

4. "जगी सर्व सुखी असा कोण आहे?" (*ɟəgiː sərvə sukʰiː əsaː koːɳə aːɦeː?)(Who is happy/satisfied in the world?)*

   In this sentence जगी (ɟəgiː) is the short inflected form of post position added verb.

- As compared to English morphology of Marathi language is extra difficult.
- Proper noun identification is an open problem in Marathi language as unlike English it does not have a system of capital case and small case letters.
- Non availability of standard electronic resources such as Marathi corpus with proper distribution of training and testing documents makes it difficult to compare Marathi text classification systems.

These issues impose constraints and hurdles in developing automatic Marathi text classification system.

IV. EXPERIMENTAL SETUP

This section presents the experimental setup used for experimenting Marathi text document classification namely datasets, classification methods and evaluation metrics.

A. Classification Methods

The literature on the subject has prescribed different classifiers for text classifications.

Out of the many classifiers we have chosen Naïve Bayes (NB), Centroid Based (CB), K-Nearest Neighbor (KNN) and Modified KNN (MKNN) for our implementation and for testing purpose. Detailed discussion on these methods can be found in [25]. The entire implementation work is done in Java.

Table-II: Distribution of our Sport Dataset

Classification Category	No. of Documents	Total Sentences	Total Words	Words after *Stopword* Removal	Total Unique words
Badminton	100	1656	16317	10365	3084
Cricket	100	1855	22301	14439	4901
Football	100	1915	21872	14262	4452
Total	300	5426	60490	39066	12437

The implementation of the centroid-based methods is based on the implementation of the Vector method. For the KNN method Cosine similarity is used to find similarity of testing document with each training document, and only the 5 nearest documents are considered i.e. K is set to 5. Selection of distance/similarity measure is done by performing experiments by using different similarity (Dice, Inner, Jaccard, Cosine) or distance (Euclidean, Minkoswski and Manhattan) measures. MKNN algorithm is implemented, various experiments are conducted for selection of H values, H=21 is selected and for factor α, α=0.5 has been used. The normalized *tf-idf* term weighting approach has been applied.

B. Evaluation Metrics

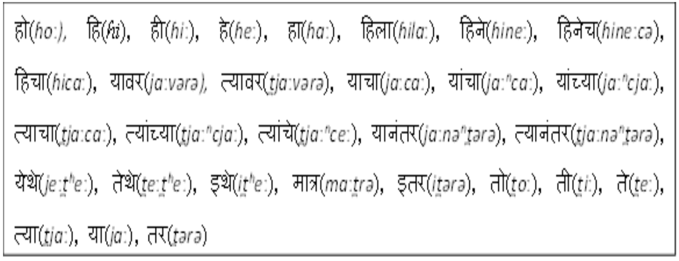
Text classification literature advocates evaluation metrics such as Precession, Recall and F-Score. These metrics have been used to test performance of individual category. Other sub measures such as Micro and Macro Average of Precision, Recall and F-Score have been used for measuring the performance of over all categories. Discussion on these measures in the context of text classification is found in [26].

C. Datasets

▪ Creating a New Sport Dataset from News in e-Sakal Marathi News Paper

The corpus used for Marathi text classifications was created from the online sports related articles published during March 2013 to Dec. 2014 from popular Marathi newspaper e-sakal. These sports related documents were belongs to three categories viz Badminton, Cricket and Football. Downloaded documents were cleaned and converted into plain .txt files for further processing. Table-II summarizes our sport dataset.

For experimentation purpose, we have prepared our own stopwords list. Stopwords are generally the function words i.e. closed POS category words of a natural language and does not contributes to classification. Beside these function words we have also considered the words which occur with more than 80% of documents [28] as stopwords. Few sample stopwords are given below.


हो(*ho:)*, हि(*ɦi*), ही(*hi:*), हे(*he:*), हा(*ha:*), हिला(*hila:*), हिने(*hine:*), हिनेच(*hine:cə*), हिचा(*hica:*), यावर(*ja:vərə),* त्यावर(*t̪ja:vərə*), याचा(*ja:ca:*), यांचा(*ja:"ca:*), यांच्या(*ja:"cja:*), त्याचा(*t̪ja:ca:*), त्यांच्या(*t̪ja:"cja:*), त्यांचे(*t̪ja:"ce:*), यानंतर(*ja:nə"t̪ərə*), त्यानंतर(*t̪ja:nə"t̪ərə*), येथे(*je:t̪ʰe:*), तेथे(*t̪e:t̪ʰe:*), इथे(*it̪ʰe:*), मात्र(*ma:t̪rə*), इतर(*it̪ərə*), तो(*t̪o:*), ती(*t̪i:*), ते(*t̪e:*), त्या(*t̪ja:*), या(*ja:*), तर(*t̪ərə*)

It was observed that the dataset unique words reduced to12347 unique words

For our experimentation purpose dataset documents were split into two groups as training and testing documents. The ratio chosen was 80:20 (80 training and 20 testing for each category viz. Badminton, Cricket and Football). These documents were trained and tested on Naïve Bayes, Centroid Based, K-Nearest Neighbor and Modified KNN.

Table-III: Distribution of Training and Testing Documents

Category	**Training Documents**	**Testing Documents**	**Total**
Badminton	80	20	100
Cricket	80	20	100
Football	80	20	100
Total	240	60	300

For comprehensive experimentation purpose we have generated 10 random samples, each comprising of 240 training documents and 60 testing documents from all categories (i.e. 80 training and 20 testing documents from each category). Detailed summary statistics of training documents is presented in Table IV below

Table-IV: Resulting Vocabulary (Unique words) of Training Dataset after removal of Stop words

Statistic of Training Dataset			
Sample No.	*Total Words*	*Total words after Stopword Removal*	*Unique words*
Sample1 (all categories)	48988	31386	8953
Sample2	47692	30867	8840
Sample3	49781	31998	9085
Sample4	48578	31351	9003
Sample5	48762	31549	9031
Sample6	49509	31845	9030
Sample7	48839	31653	8960
Sample8	48521	31212	8930
Sample9	48392	31314	8893
Sample10	48469	31205	8940
Average	48753.1	31438	8966.5

D. Text Pre-processing

Pre-processing was done to reduce the feature space which also makes data more uniform and increases the performance.

Following preprocessing steps were followed

- UTF-8 encoding conversion of dataset.
- Tokenization: removing extra spaces, hyphen, punctuation marks, numbers, digits and non-Marathi alphabets letters.
- Stopword removal.
- Vector Space Model representation of Marathi text documents.

During tokenization we face the problem due to hyphenation and full stop (.) characters in dataset documents.

For examples आय-लीग (ɑːjə-liːɡə), के. श्रीकांत (keː.ɕriːkɑːⁿt̪ə)

In above examples words should be tokenized as आय (ɑːjə), लीग (liːɡə), के (keː) and श्रीकांत (ɕriːkɑːⁿt̪ə).

V. Results and Discussion

We have applied Naïve Bayes, Centroid Based, K- Nearest Neighbor and Modified KNN Classifiers to all 10 randomly generated samples. Categorywise 10 samples classification results for each classifier are presented below (Table V, Table VI, Table VII and Table VIII respectively)

A. Results of Naïve Bayes Classifier:

Table-V: Category wise ten sample's results with NB classifier

Sample No.	Badminton			Cricket			Football		
	Precision (%)	*Recall (%)*	*F1- Score (%)*	*Precision (%)*	*Recall (%)*	*F1- Score (%)*	*Precision (%)*	*Recall (%)*	*F1- Score (%)*
Sample1	100	100	100	95.24	100	97.56	100	95	97.44
Sample2	100	100	100	100	100	100	100	100	100
Sample3	100	100	100	95.24	100	97.56	100	95	97.44
Sample4	100	100	100	100	100	100	100	100	100
Sample5	100	95	97.44	95.24	100	97.56	95	95	95
Sample6	100	100	100	100	100	100	100	100	100
Sample7	100	100	100	100	100	100	100	100	100
Sample8	100	100	100	100	100	100	100	100	100
Sample9	100	95	97.44	100	100	100	95.24	100	97.56
Sample10	100	100	100	100	100	100	100	100	100
Average	100	99	99.488	98.572	100	99.268	99.024	98.5	98.744

B. Results of Centroid Based Classifier

Table VI: Category wise ten samples results with Centroid Based classifier

Sample No.	Badminton			Cricket			Football		
	Precision (%)	*Recall (%)*	*F1-Score (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Sample1	100	100	100	95.24	100	97.56	100	95	97.44
Sample2	100	100	100	100	100	100	100	100	100
Sample3	100	100	100	95.24	100	97.56	100	95	97.44
Sample4	100	100	100	100	100	100	100	100	100
Sample5	100	95	97.44	95.24	100	97.56	95	95	95
Sample6	100	100	100	100	100	100	100	100	100
Sample7	100	100	100	100	100	100	100	100	100
Sample8	100	100	100	100	100	100	100	100	100
Sample9	100	95	97.44	100	100	100	95.24	100	97.56
Sample10	100	100	100	100	100	100	100	100	100
Average	100	99	99.49	98.57	100	99.27	99.02	98.5	98.74

C. Results of KNN Classifier

Deciding the value of K is most critical issue in any KNN experimentation. Following section describes the same in our experimentation. Accuracy of K-NN is also depends on choice of distance/similarity measures used in classification phase.

KNN has been applied to text categorization since the early days of its research [29] it is known to be one of the most effective methods on a benchmark corpus i.e. Reuter's corpus of Newswire stories [30].

- *K Parameter for KNN Classifier*

Selecting appropriate value for K and choice of distance/similarity measure used in classification phase plays an important role in KNN classifier. Deciding the value of K is most critical issue in any KNN experimentation. Following section describes the same in our experimentation. Accuracy and success of KNN very much depend on this value [30].

For selecting the value of K instead of using any heuristic or any other method we preferred to test this classifier for different values of K (K=3,5,7,9,11,13) on sample1. After experimentation we observed that the values 5,9,11 and 13 for K yielded same best performance of 98.33% Micro Average of F1-Score and 98.62% Macro Average of F1-Score. It was also observed that the performance saturates at K = 9, 11 and 13 (please refer Fig.1. below).

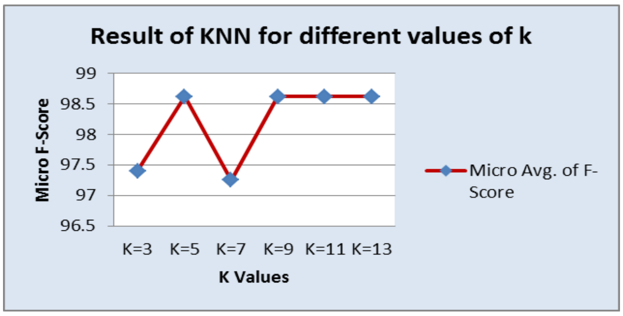


Fig. 1. Results of KNN for Various Values of K

Thus we arrived at the value of K=5 on remaining samples and calculated other performance metrics which are summarizing in Table VII.

Table VII: Category wise ten samples results with K Nearest Neighbor classifier

Sample No.	Badminton			Cricket			Football		
	Precision (%)	*Recall (%)*	*F-Score (%)*	*Precision (%)*	*Recall (%)*	*F-Score (%)*	*Precision (%)*	*Recall (%)*	*F-Score (%)*
Sample1	100	95	97.44	95	95	95	90.48	95	92.68
Sample2	100	95	97.44	95.24	100	97.56	95	95	95
Sample3	100	100	100	90.91	100	95.24	100	90	94.74
Sample4	100	100	100	95.24	100	97.56	100	95	97.44
Sample5	100	100	100	95.24	100	97.56	100	95	97.44
Sample6	100	95	97.44	95.24	100	97.56	95	95	95
Sample7	95.24	100	97.56	100	95	97.44	95	95	95
Sample8	100	100	100	90.91	100	95.24	100	90	94.74
Sample9	100	95	97.44	100	100	100	95.24	100	97.56
Sample10	100	95	97.44	100	100	100	95.24	100	97.56
Average	99.52	97.5	98.47	95.78	99	97.32	96.6	95	95.72

D. Experiment with MKNN Classifier

MKNN is an improved version of KNN. The main idea here is to find stability of all documents in the training set. It introduces new value *validity*. Validity is calculated for every document in training set. Validity value indicates stability of the training set document. Validity of the training set document is calculated using H number of nearest neighbor of each training set documents. Validity of every document in training set is calculated according to the following formula

$$validity\ (x) = \frac{1}{H}\sum_{i=1}^{H} S(label(x), label(v_i(x)))$$

If training document and its ith neighbor in the set of training documents have same label then function S(a,b) returns 1 otherwise function S(a,b) returns zero. Thus validity function gives average number of stable neighbors of training set documents. According to validity function more stable document in training set gives more validity value.

Selection of proper value of H will affect the performance of classifier. Value of H has been selected through experimentation. Initially system has been tested for various values of H (H = 9, H=13, H=17, H=21). Fig. 2. depict the performance of MKNN for various values of H. From Fig. 2. it is clear that when value of H has been changed from 13 to 17 performance of MKNN is improved from 97.25% Macro Average of F-Score to 98.62% Macro Average of F-Score, but it remains unchanged for H=21. So H=21 has been considered throughout the experimentation of MKNN.

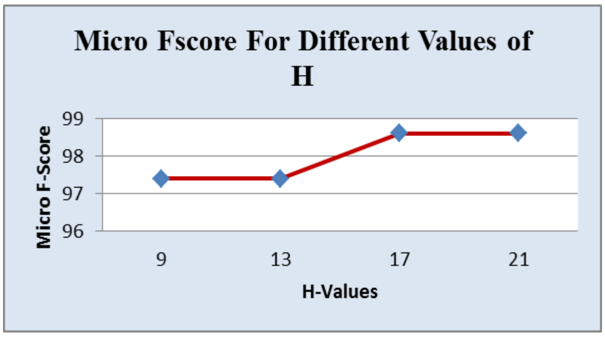


Fig. 2. Performance of MKNN for different values of H in terms of Macro Average of F-Score.

MKNN is the type of weighted KNN. Classification method of MKNN is similar to KNN classifier. In classification stage, weighted similarity of a testing document with every document in the training set is calculated. Weighted similarity of testing sample with ith training document is calculated using formula as given below,

$$W(d_i, X) = validity(x) * [Sim(X, D_i) + \alpha]$$

Value of α is set to 0.5. In MKNN top K neighbors based on stability or validity is selected. Among the top K neighbors, a test document is assigned to the category which have majority of its nearest neighbors. In this experiment, we applied MKNN algorithm to the same training data set (refer Table IV) and the results of MKNN classifier are given below (Table VIII)

Table VIII: Category wise ten samples results with Modified K Nearest Neighbor classifier

Sample No.	Badminton			Cricket			Football		
	Precision (%)	*Recall (%)*	*F-Score (%)*	*Precision (%)*	*Recall (%)*	*F-Score (%)*	*Precision (%)*	*Recall (%)*	*F-Score (%)*
Sample1	100	95	97.44	95	95	95	90.48	95	92.68
Sample2	100	95	97.44	95.24	100	97.56	95	95	95
Sample3	100	100	100	90.91	100	95.24	100	90	94.74
Sample4	100	100	100	90.91	100	95.24	100	90	94.74
Sample5	100	100	100	90.91	100	95.24	100	90	94.74
Sample6	100	100	100	95.24	100	97.56	100	95	97.44
Sample7	100	100	100	95.24	100	97.56	100	95	97.44
Sample8	100	100	100	95.24	100	97.56	100	95	97.44
Sample9	100	94.75	97.31	95.24	100	97.56	95	95	95
Sample10	100	95	97.44	95.24	100	97.56	95	95	95
Average	100	97.975	98.963	93.917	99.5	96.608	97.548	93.5	95.422

Table IX shows category wise average of all ten samples for Naïve Bayes, Centroid Based, K Nearest Neighbor and MKNN classifiers. It also represents average of Precision, Recall and F-Score for all three categories using Naïve Bayes, Centroid Based, K Nearest Neighbor and MKNN classifiers.

Table IX: Category wise ten sample's average for Naïve Bayes, Centroid Based, K Nearest Neighbor and MKNN classifiers.

Category	Naïve Bayes			Centroid Based			K Nearest Neighbor			Modified KNN		
	Precision (%)	*Recall (%)*	*F1-Score (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Badminton	**100**	99	99.488	**100**	99	99.49	**99.52**	97.5	98.47	**100**	97.98	98.963
Cricket	98.572	**100**	99.268	98.57	**100**	99.27	95.78	**99**	97.32	93.917	**99.5**	96.608
Football	**99.024**	98.5	98.744	**99.02**	98.5	98.74	**96.6**	95	95.72	**97.548**	93.5	95.422
Average	**99.1987**	99.17	99.167	**99.1967**	99.167	99.167	**97.3**	97.17	97.17	**97.155**	96.99	96.998

Fig.3 shows classifier wise F1-Score for all four classifiers. From Fig.3 it is cleared that Naïve Bayes and Centroid Based gives best performance among all four classifiers with 99.167% F1-Score followed by KNN having 97.17% F1-Score and MKNN having lowest performance with 96.998% F1-Score.

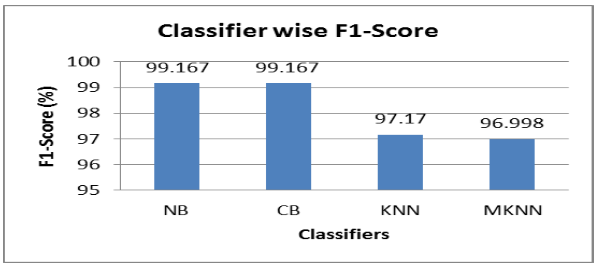


Fig. 3. F1-Score Percentage of all four Classifiers

We have also calculated dMicro and Macro averages F1-Score which are summarized in Table X.

Table X: Micro and Macro Averages Score for NB, CB, KNN and MKNN

Classifier	Micro Average Precision (%)	Micro Average Recall (%)	Micro Average F1-Score (%)	Macro Average Precision (%)	Macro Average Recall (%)	Macro Average F1-Score (%)
NB	99.166	99.166	99.166	99.198	99.166	99.166
CB	99.166	99.166	99.166	99.198	99.166	99.17
KNN	97.167	97.167	97.167	97.277	97.167	97.7
MKNN	97.161	97.161	97.161	97.321	97.158	96.997

VI. Conclusion:

In this paper we have performed experimentations on Automatic Marathi text classification for the newspaper dataset. This study reports some of the issues in the context of Marathi language text classification. Morphological nature of Marathi language text is also presented.

For our experimentation of automatic Marathi text document classification, the dataset used was created from the online sports related 300 articles published in popular Marathi newspaper e-sakal. These sports related documents were belongs to three categories viz Badminton, Cricket and Football. We have performed our experimentation by randomly selecting 240 documents for training and 60 documents for testing from total 300 documents. We have generated 10 different samples by repeating the same procedure 10 times.

Different classifiers such as Naïve Bayes, Centroid Based, K-Nearest Neighbor and Modified KNN were implemented and applied on 10 different samples. Average of these 10 sample's result was considered as classification result for particular classifier. The overall study concludes that all the four classifier's results are acceptable for Marathi Language text. Among four classifiers, Naïve Bayes and Centroid Based give best performance with 99.166% Micro and Macro Average of F-score and MKNN gives lowest performance with 97.16% Micro Average of F-Score and 96.997% Macro Average of F-score. We observed that the classification speed of NB is very fast among all the four classifiers. As KNN and MKNN find similarity of each test document with every training document at the time of testing, KNN and MKNN required more time to classify test document. As MKNN required one time pre-processing it takes more time than KNN in training phase. This study deals with classification without feature selection. In future we will study the effect of feature selection on classification using same classification methods.

REFERENCES

1. Available online at https://en.wikipedia.org/wiki/ Marathi_language. Last visited on 02-06-2017
2. Arts, South Asian. "Encyclopædia Britannica. Encyclopædia Britannica 2007 Ultimate Reference Suite.
3. Dhoṅgaḍe, Rameśa; Wali, Kashi (2009). "Marathi". London Oriental and African language library. John Benjamins Publishing Company. 13: 101, 139. ISBN 9789027238139.
4. Maharashtra tops list of Internet subscribers in India | Business Line. Available online at http://www.thehindubusinessline.com /info-tech/maharashtra-tops-list-of-Internet-subscribers-in-india/article9367201.ece. Nov. 20 2016. Last visited on 22-03-2017.
5. Zakaria Elberrichi and Karima Abidi, " Arabic Text Categorization: A comparative Study of Different Representation Modes", The International Arab Journal of Information Technology Vol. 9 No. 5 September 2012.
6. Eibe Frank and Remco R. Bouckaert., "Naive Bayes for text classification with unbalanced classes". In: Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases, 2006, Springer, Berlin. pp. 503-510.
7. Masud Karim, Rashedur M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing", Journal of Software Engineering and Applications, 2013, 6, 196-206, Published Online April 2013
8. Fouzi Harrag, Abdul Malik Salman Al-Salman, Mohammed Ben Mohammed, "A Comparative Study of Neural Networks Architectures on Arabic Text Categorization Using Feature Extraction", 978-1-4244-8611-3/10/$26.00 ©2010 IEEE
9. Mansur, Munirul. "Analysis of n-gram based text categorization for bangla in a newspaper corpus." PhD diss., BRAC University, 2006.
10. Vikramjit Mitra, Chia-Jiu Wang, Satarupa Banerjee, "Text classification: A least square support vector machine approach", Science direct, Applied Soft Computing 7 (2007) 908–914, doi:10.1016/j.asoc.2006.04.002
11. Eui-Hong(Sam) Han and George Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results", Principles of Data Mining and Knowledge Discovery, p 424-431, 2000.
12. Johnson, David E., Frank J. Oles, Tong Zhang, and Thilo Goetz. "A decision-tree-based symbolic rule induction system for text categorization." IBM Systems Journal 41, no. 3 (2002): 428-437.
13. Patil, Ajay S., and B. V. Pawar. "Automated classification of web sites using Naive Bayesian algorithm." In Proceedings of the international multiconference of engineers and computer scientists, vol. 1. 2012.
14. Rupali P. Patil, R. P. Bhavsar, B. V. Pawar, "A Note on Indian Languages Text Classification", Asian Journal of Mathematics and Computer Research, Vol. 15 Issue 4, Jan 2017, 41-55.
15. Rajnish M. Rahkolia and Jitendrakumar R. Saini, 'Classification of Gujarati Documents using Naïve Bayes Classifier", Indian Journal of Science and Technology, Vol.10(5), February 2017, DOI:10.17485/ijst2017/v1015/103233.
16. Rajan, K., Vennila Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan. "Automatic classification of Tamil documents using vector space model and artificial neural network." Expert Systems with Applications 36, no. 8 (2009): 10914-10918.
17. M NarayanaSwamy, M. Hanumanthappa, "Indian Language Text Representation and Categorization Using Supervised Learning Algorithm", International Journal of Data Mining Techniques and Applications, Vol.: 02, December2013, pages 251-257.
18. Mandal, Ashis Kumar, and Rikta Sen. "Supervised Learning Methods for Bangla Web Document Categorization." International Journal of Artificial Intelligence & Applications 5, no. 5 (2014): 93.
19. Mrs. Sushma R. Vispute, Prof. M. A. Potey, "Automatic Text Categorization of Marathi Documents Using Clustering Technique", Advanced computing Technologies (ICACT), 2013 15th International IEEE Conference,21-22 Sept 2013, pages1-5, 978-1-4673-2816-6, DOI:10.1109/ICACT.2013.6710543.
20. Patil Meera, and Pravin Game. "Comparison of Marathi Text Classifiers." International Journal on Information Technology 4, no. 1 (2014): 11.
21. Patil Jaydeep Jalindar, and Nagaraju Bogiri. "Automatic text categorization: Marathi documents." In Energy Systems and Applications, 2015 International Conference on, pp. 689-694. IEEE, 2015.
22. N. Dangre, A. Bodke, A. Date, S. Rungta, S. S. Pathak, "System for Marthi News Clustering", 2nd International Conference on Intelligent Computing, Communication and Convergence (ICCC-2016) Procedia Computer Science 92(20160 18-22, published by Elsevier B.V.
23. Aishwarya Sahani, Kaustubh Sarang, Sushmita Umredkar and Mihir Patil, "Automatic Text Categorization of Marathi Language Documents", International Journal of Computer Science and Information Technologies, Vol. 7(5), 2016, 2297-2301. ISSN: 0975-9646.
24. Pooja Bolaj and Sharvari Govilkar, "Text Classification for Marathi Documents using Supervised Learning Methods", International Journal of Innovation and Advancement in Computer Science (IJIACS), Volume 6, Issue 8, August 2017. ISSN 2347-8616.
25. Rupali Patil, Bhavsar R.P., Pawar B.V., "Holy grail of hybrid text classification", International Journal of Computer Science Issues; Volume 13, Issue 3, May 2016, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.
26. Rupali P. Patil, Bhavsar R. P, Pawar B. V., "A comparative study of text classification methods: An experimental approach." International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC). 2016; 4(3):517 – 523. ISSN: 2321-8169.
27. E-Sakal, Online at http://www.esakal.com Last visited on 15-08-17.
28. Savoy, Jacques. "Searching strategies for the Bulgarian language." information retrieval 10, no. 6 (2007 Dec. 1): 509-529.
29. Sebastiani, Fabrizio. "Machine learning in automated text categorization"ACM computing surveys (CSUR) 34, no.1(2002): 1
30. Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "An KNN model-based approach and its application in text categorization." In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 559-570. Springer Berlin Heidelberg, 2004.

AUTHORS PROFILE

Dr. Rupali P. Patil obtained her B.Sc. degree from Amaravati University (1995) followed by M.Sc. (Computer Science) from Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon in 1997. She was awarded Ph.D. in Computer Science in 2018 from the same university. She has 10 years of teaching experience in the field of Computer Science. Presently she is working as Assistant Professor at Department of Computer Science, S.S.V.P.S's L.K. Dr. P.R.Ghogrey Science College Dhule, Maharashtra. Her research areas include NLP, Machine Learning and Information Retrieval. She is member of ACM-W.

Dr. R. P. Bhavsar obtained his B.C.S. degree from Pune

University (1992) followed by MCA in 1995) from Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. He was awarded Ph.D. in Computer Science in 2016 from the same university. He has varied experience of 23 years which includes 03 years as Scientist at C-DAC (Pune), 07 years as System Analyst, 13 years of teaching. Currently, he is working as Associate Professor in School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. His research area includes Machine Translation, Applied Natural Language Processing, Capacity Building, Machine Learning, IT & Networking infrastructure setup and has handled software development projects. Presently, 3 research scholars are pursuing their Ph. D. under his supervision. He is a life time member of Computer Society of India (CSI)

Prof. B. V. Pawar finished his B. E. (Production Engineering) from reputed VJTI Mumbai (1986), followed by M.Sc. (Computer science) from Mumbai University (1988). He was awarded Ph.D. degree in Computer Science by Kavayitri Bahinabai Chaudhari North Maharashtra University (formally North Maharashtra University), Jalgaon, India. He has huge experience of 30 years teaching and research in the field of computer science. Presently, he is working as Professor at School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. His research interests include NLP, Web computing, Information Retrieval and Machine Translation, Machine Learning. He has guided 12 research Scholars for their doctoral research. He has published over 75 research /journal papers and 113 conference papers presentations and 05 invited talks. He is a life time member of Computer Society of India (CSI), Linguistic Society of India (LSI) and International Association of Engineers (IAENG).