**Indian Journal of Science and Technology,** *Vol 10(5), DOI: 10.17485/ijst/2017/v10i5/103233, February 2017*
*ISSN (Print) : 0974-6846*
*ISSN (Online) : 0974-5645*

# Classification of Gujarati Documents using Naïve Bayes Classifier

**Rajnish M. Rakholia[1*] and Jatinderkumar R. Saini[2]**

[1]School of Computer Science, R. K. University, Rajkot - 360020, Gujarat, India; rajnish.rakholia@gmail.com
[2]Narmada College of Computer Application, Bharuch - 392011, Gujarat, India; saini_expert@yahoo.com

## Abstract

**Objectives:** Information overload on the web is a major problem faced by institutions and businesses today. Sorting out some useful documents from the web which is written in Indian language is a challenging task due to its morphological variance and language barrier. As on date, there is no document classifier available for Gujarati language. **Methods:** Keyword search is a one of the way to retrieve the meaningful document from the web, but it doesn't discriminate by context. In this paper we have presented the Naïve Bayes (NB) statistical machine learning algorithm for classification of Gujarati documents. Six pre-defined categories sports, health, entertainment, business, astrology and spiritual are used for this work. A corpus of 280 Gujarat documents for each category is used for training and testing purpose of the categorizer. We have used k-fold cross validation to evaluate the performance of Naïve Bayes classifier. **Findings:** The experimental results show that the accuracy of NB classifier without and using features selection was 75.74% and 88.96% respectively. These results prove that the NB classifier contribute effectively in Gujarati documents classification. **Applications:** Proposed research work is very useful to implement the functionality of directory search in many web portals to sort useful documents and many Information Retrieval (IR) applications.

**Keywords:** Classification, Document Categorization, Gujarati Language, Naïve Bayes

## 1. Introduction

To retrieve the relevant documents from the web is a significant task to satisfy the demands of different users. It is more difficult for the resource poor language like Gujarati, Panjabi, Marathi and other Indian languages. Manual document classification is time consuming process, which makes it infeasible for handling the huge number of text documents[1]. Automatic document classification is a one of the way to cope such a type of problem to save human efforts and increase the speed of the system. Six predefined categories (sports, health, entertainment, business, astrology and spiritual) are used for this work. Main objective of this research is to enhance the performance of Information Retrieval (IR) and other Natural Language Processing (NLP) applications such as library system, mail classification, spam filtering, sentiment analysis and survey classification etc., for Gujarati language. In proposed work, Naïve Bayes (NB) classifier is used. Basics of Gujarati language and machine learning approach are as follows:

### 1.1 Gujarati Language

Gujarati is an official and regional language of Gujarat state in India. It is 23rd most widely spoken language in the world today, which is spoken by more than 46

*Author for correspondence

million people. Approximately 45.5 million people speak Gujarati language in India and half million speakers are from outside of India that includes Tanzania, Uganda, Pakistan, Kenya and Zambia. Gujarati language is belongs to Indo-Aryan language of Indo-European language family and it is also closely related to Indian Hindi language.

### 1.2 Naïve Bayes (Supervised Machine Learning Algorithm)

Naïve Bayes (NB) is a most popular statistical machine learning algorithm for text classification. In regards to the existing algorithms, Naïve Bayes algorithm is potentially good against several approaches for document classification (such as decision tree, neural network, and support vector machines) in the terms of simplicity[2,3]. NB worked quite well in many real world applications such as document and text classification, but small amount of training is needed to estimate the required parameters.

### 1.3 Document Classification

Document classification is an important task in information science and library science. In this task assign one or more label, class or category to each document. Manually category assignment is a better approach in library science when less number of documents is present. But in information science algorithmically approach is better due to huge amount of documents available.

### 1.4 Existing work on Indian Languages

A number of machine learning algorithms have been used for document and text categorization for Indian languages by different researchers. Table 1 gives summary and comparison of various classification algorithm, feature extraction technique and accuracy of related work on document categorization, mainly for Indian languages.

### 1.5 Existing work on Non-Indian Languages

For the movies review document classification[12] used Naïve Bayes (NB) and Support Vector Machine (SVM). Pre-defined two classes (categories) positive and negative were used to assign document labels. Unigram was also used with one of the classification technique. Researchers' in[12] performed document classification for movies review dataset using Naïve Bayes, Maximum Entropy and Support Vector Machine with n-gram model. They also been found that unigram perform well then bigram with all three machine learning technique.

Researchers' in[13] performed experiment to evaluate different feature selection methods with most popular machine learning algorithms NB, SVM, k-nearest-neighbors (kNN) and Rocchio-style classifier. X-square statistics feature selection method performed quite well than others (IG, IG2 and DF). Whereas, Author in[14] evaluated the performance and results of twelve feature selection technique to examine which works better. Based on him study, it has been found that IG (Information Gain) worked better than other techniques.

Five machine learning algorithms and four feature selection techniques for the Chinese document classification. Based on their experiment, it has been found that Information Gain (IG) and Support Vector Machine (SVM) produced better result than other feature selection technique and machine learning algorithm respectively[15]. Hybrid classification approach (combined machine learning algorithm + rule-based classification), and 10-fold cross validation method were used to evaluate the performance of proposed approach[16]

Naïve Bayes and support Vector Machine were used for Arabic document classification[17]. They created more than 700 documents for each category from different seven news categories. They achieved 77% and 74% accuracy for SVM and Naïve Bayes algorithm respectively. Naïve Bayes classifier and TF-IDF was used as feature selection. Total five categories were considered for data collection. They created 300 documents for each category from Arabic news website for experiment[18]. They achieved accuracy of 90%.

Based on literature review of document classification for Indian and non-Indian languages, we conclude that majority researchers have used Naïve Bayes classifier and TF-IDF for feature selection.

## 2. Naïve Bayes Classifier

This section organized as follows: Section 2.1 described preprocessing steps required for document classification, 2.2 Feature selection, 2.3 NB Training phase, 2.4 Posterior probability computation, 2.5 Dataset, 2.6 k-fold cross validation.

**Table 1.** Comparison of existing work

| Sr. No. | Author (Year) [References] | Classification Approach used | Feature Selection | Data source / Corpus | Language | Result / Accuracy |
|---|---|---|---|---|---|---|
| 1 | [4] | Label Induction Grouping Algorithm based on SVM | TF-IDF | They created own corpus of 200 document with more than 10 news categories | Marathi | They achieved efficient result for Marathi document classification. |
| **2** | [5] | Naïve Bayes | Feature extraction was performed using Marathi words dictionary | Five categories Literature, Economy, Botany, Geography and History has used for data collection. They Created more than 800 documents for each category. | Marathi | They achieved higher accuracy by using Naïve Bayes classifier where as k-NN produce least accuracy among. |
| | | Centroid Based | | | | |
| | | k-NN | | | | |
| **3** | [6] | Naïve Bayes | TF-IDF | Created own corpus of 1000 documents from various Bangla websites. They considered five categories (Business, Sports, Health, Education and Technology) | Bangla | 85.22% |
| | | SVM | | | | 89.14% |
| | | k-NN | | | | 74.22% |
| | | Decision Tree (C4.5) | | | | 80.65% |
| **4** | [7] | Naïve Bayes | TF-IDF | Created 800 documents from the web (Telugu news papers) science, economics, sports, politics, culture and health domain were used for data collection | Telugu | Results shown that, SVM performed quit well than Naïve Bayes and k-NN. |
| | | SVM | | | | |
| | | k-NN | | | | |
| 5 | [8] | Naïve Bayes | TF-IDF | South Indian language corpus(own created) 100 documents related to cinema for each language | Telugu, Kannada, Tamil | 97.66% |
| | | Decision Tree | | | | 97.33% |
| | | k-NN | | | | 93% |
| **6** | [9] | Naïve Bayes | TF-IDF | Created 180 documents from the web (Panjabi news paper) Cricket, Football, Kabbadi, Tennis, Hockey, Badminton and Olympics sports categories were used for data collection | Panjabi | 64% |
| | | Centroid Based | | | | 71% |
| | | Ontology based | | | | 85% |
| | | Hybrid based | | | | 85% |
| **7** | [10] | Artificial Neural Network | TF-IDF | Tamil CIIL corpus (CIIL, Mysore India) | Tamil | 93.33% |
| | | Vector Space Model | | | | 90.33% |
| **8** | [11] | Naïve Bayes | TF-IDF | DoE-CIIL corpora and created own corpus from Telugu news websites | Major ten Indian languages | SVM out-performed than Naïve Bayes and k-NN. |
| | | SVM | | | | |
| | | k-NN | | | | |

## 2.1 Pre-Processing Steps

Main objective of preprocessing phase in document classification is to enhance the influence between word and category of document. It is important step to discard the most insignificant and irrelevant words to improve the quality of document[19,20]. Steps of preprocessing for document classification as follows:

### 2.1.1 Tokenization

It is a process to divide texts into number of individual tokens to reduce the unnecessary contents from the document. JAVA utility package and space delimiter were used to done this process. All special characters and punctuation mark have also been removed in this step.

### 2.1.2 Stop Words Elimination

Till now, there is no unique stop words list is available for Indian Gujarati language. With the help of linguistic experts and by manual inspection, we have manually constructed a list of 531 stop words. This stop words list is only domain specific that includes sports, entertainment, health, business, spiritual and astrology.

### 2.1.3 Stemming

For the Gujarati language, there is no automation tool is available to create stemmed words list from dataset or corpus. We have used hand crafted Gujarati suffix list in order create a list of stemmed words[21–26].

## 2.2 Feature Selection

It is the process of selecting most relevant key words from the document based on its frequency and contribution (weight) in the document. In this research, we have used TF-IDF feature selection technique.

TF-IDF (Term Frequency-Inverse Document Frequency) weight is a statistical measure which is used to evaluate; how particular word is important for the document from collected dataset or a corpus. Computing functions of TF and IDF are as follows.

### 2.2.1 TF (Term Frequency)

Which measure; how frequently a word occurred in a particular document. Frequency of the word is also based on length of the document. Long document may contain more occurrences of the word than short document. In TF calculation all terms to be considered equal importance. TF could be calculated using following formula:

$$\text{TF (term)} = \frac{\text{occurrences of particular term in a document}}{\text{total number of terms in whole document}} \quad (1)$$

### 2.2.2 IDF (Inverse Document Frequency)

Which measure; how particular term is important for the document. IDF could be calculated as logarithm of number of documents in whole corpus divided by number of document contained particular term. IDF could be calculated using following formula:

$$\text{IDF (term)} = \log_e\left(\frac{\text{total number of documents in whole corpus}}{\text{number of document that contains a term}}\right) \quad (2)$$

Product of Equation (1) and (2) is used to produce a composite weight for each term in each document (TF*IDF).

## 2.3 NB Training Phase

This learning phase is based on Naïve Bayes algorithm. Main task of this phase is to assign a label to the newly encountered document from the pre-defined categories $c_j$.

Let C be the set of pre-defined classification categories and D be the set of labeled documents.

Each category $c_i$ $C$

W is a set of distinct words in whole training set. $w_k$ W, where $w_k$ is a word/term.

Each document $d_i$ $D$ and $d_i = \{w_1, w_2, w_3, \ldots, w_n\}$,

Let $D_i$ be the subset of D in the category of $c_i$.

Compute the probability of each category $c_i$ :

$$P(cj)=|Dj|/|D| \quad (3)$$

Where,

$|D_j|$ = the number of documents in the category $c_j$.

$|D|$ = the total number of documents in all categories.

Compute the probability of each word/term in category $c_j$.

$$P(wk/cj)= (Ok,j +1)/(sj +| ATj |) \quad (4)$$

Where,

$O_{k,j}$ = the number of times $w_k$ occurs in category $c_j$.

$s_j$ = total number of distinct words in category $c_j$.

$|AT_j|$ = total number of words in single document which is created by merging all documents of category cj.

Two Naïve assumptions are considered in learning phase for Gujarati language: Sequence of the words in the document does not affect the document classification. In the same document, probability of specific encountered word is the same regardless the position of the word.

$$P(d_i|c_j)=P(\{w_1,w_2,w_3,\ldots\ldots,w_n\}|c_j) \quad (5)$$

Where, $w_1, w_2, w_3,\ldots\ldots,w_n$ W

The probability of occurrence of particular word in a document is independent of the occurrence of other words.

$$P(\{w_1,w_2,w_3,\ldots\ldots,w_n\}|c_j)=P(w_1|c_j)*P(w_2|c_j)*\ldots*P(w_n|c_j) \quad (6)$$

**Posterior Probability Computation**

$$P(cj|di) = \frac{\left[P\left(c_j\right) * P\left(d_i|c_j\right)\right]}{P(d_1)} \quad (7)$$

Where,

$$P(d_i)=P(c_1)P(d_i|c_1) + P(c_2)P(d_i|c_2)+ P(c_3)P(d_i|c_3)+\ldots\ldots+ P(c_6)P(d_i|c_6)$$

The document $d_ic_k$, if maximum $\{P(c_j|d_i), j=1,2,3,4,5,6\}= P(c_k|d_i)$

## 2.4 Dataset

We have considered six different domain specific categories (Sports, Health, Business, Entertainment, Spiritual and Astrology) for proposed research. 280 web documents were collected for each category from various Gujarati news website. Domain specific documents bifurcated by the news website itself.

## 2.5 k-fold Cross Validation

*k*-fold cross validation is a most common technique to evaluate the performance of the classifier. We have applied this technique to estimate the performance of Naïve Bayes classifier for Gujarati documents classification. At the initial stage, dataset is partitioned into k fold $F_1$, $F_2$, $F_3$,……,$F_k$. with equal size (approx). Value of k is depending on the size of dataset. For each of the k experiment, k-1 folds are used to train the model and remaining for testing. For the each iteration k-1 folds are used for the training which will be test on remaining one fold. For instance, for single iteration $F_1$, $F_2$, $F_4$,……..,$F_k$ folds are used for the training and $F_3$ will be used for testing purpose.

1. Let m be the entire dataset
   *[1680 documents including six different categories]*
2. Arrange documents of entire dataset in random order.
3. Divide dataset into *k*-folds (k chunks: m/k).
4. For i=1,….,k
   a. Train the NB using entire dataset except fold-i.
      *[For the first iteration $F_2,F_3,F_4,\ldots,F_k$ folds will be used in training]*
      (Compute this step using NB learning- Training phase)
   b. Test the NB using all the documents in fold-i.
      *[For the first iteration $F_1$ fold will be used for testing]*
   c. Compute $wc_i$, wrongly classified document from the testing.
      *[From step-b]*
5. Compute the error rate of classifier

$$E = \frac{\sum_{i=1}^{k} wc_i}{m}$$

Return to step 4, to execute next fold. To obtain accurate estimation of the classifier, *k*-fold cross validation was run multiple times by changing the sequence of documents [in step-2].

# 3. Results of NB Classifier (Without Features Selection)

For the training and testing, total number of documents in corpus is 1680. To evaluate the performance (results) of NB classifier for Gujarati language we have used *k-fold cross validation*. Experimental results and error rate of NB classifier using 10-fold, 8-fold, 6-fold, 4-fold and 2-fold cross validation has described in Table 2, 3, 4, 5 and 6 respectively. To prepare confusion matrix for each fold we have considered single randomly partition. Each confusion matrix contained classifier predicted label (columns), document belong from actually category (rows) and error rate of classifier.

**10-fold cross validation [Testing]:** 10-fold means total corpus (1680 documents) will be divided into 10 equal partitions, and each partition that contained 168 documents (Sports-32, Business-40, Entertainment-28, health-18, Spiritual-24 and Astrology-26). Table 2 presents error rate of Naïve Bayes classifier for randomly selected partition using 10-fold cross validation method.

**8-fold cross validation:** Each fold contained 210 documents (Sports-38, Business-36, Entertainment-37, health-42, Spiritual-21 and Astrology-36). Table 3 presents error rate of Naïve Bayes classifier for randomly selected partition using 8-fold cross validation method.

**6-fold cross validation:** Each fold contained 280 documents (Sports-40, Business-49, Entertainment-50, health-53, Spiritual-47 and Astrology-41). Table 4 presents error rate of Naïve Bayes classifier for randomly selected partition using 6-fold cross validation method.

**4-fold cross validation:** Each fold contained 420 documents (Sports-74, Business-71, Entertainment-63, health-81, Spiritual-68 and Astrology-63). Table 5 presents error rate of Naïve Bayes classifier for randomly selected partition using 4-fold cross validation method.

**2-fold cross validation:** Each fold contained 840 documents (Sports-164, Business-134, Entertainment-147, health-161, Spiritual-121 and Astrology-113). Table 6 presents error rate of Naïve Bayes classifier for randomly selected partition using 2-fold cross validation method.

Based on obtained results for NB without feature selection we conclude that, accuracy obtained in 10-fold cross validation is quite well then accuracy of 2-folds. Because in 2-fold cross validation whole data set is divided into two portions, one portion is used for training and other for testing. Due to that reason, documents included in testing it may not be properly trained by the classifier.

## 4. Results of NB Classifier (using Features Selection)

To evaluate the performance of NB classifier using features selection , experiments were done by considering 20, 40, 60,…,1000 terms which are best represent the six predefined categories. We have repeated the experiment for each number of terms using different k-folds. A summary of experiment is presented in Table 7 which is average accuracy of all predefined category for particular number of terms.

**Table 2.** 10-fold cross validation

| NB classifier predicted | | | | | | | | Error rate |
|---|---|---|---|---|---|---|---|---|
| Actual Category belongs | Category | Sports | Business | Entertainment | Health | Spiritual | Astrology | |
| | Sports | 24 | 2 | 1 | 3 | 1 | 1 | 25.00% |
| | Business | 3 | 29 | 3 | 2 | 1 | 2 | 27.50% |
| | Entertainment | 1 | 2 | 22 | 2 | 1 | 0 | 21.42% |
| | Health | 2 | 1 | 0 | 13 | 1 | 1 | 27.77% |
| | Spiritual | 1 | 1 | 0 | 1 | 19 | 2 | 20.83% |
| | Astrology | 0 | 2 | 0 | 2 | 2 | 20 | 23.07% |
| **Average error rate for single experiment using 10-fold cross validation** | | | | | | | | 24.26% |

**Table 3.** 8-fold cross validation

| NB classifier predicted | | | | | | | | Error rate |
|---|---|---|---|---|---|---|---|---|
| Actual Category belongs | Category | Sports | Business | Entertainment | Health | Spiritual | Astrology | |
| | Sports | 27 | 02 | 02 | 05 | 01 | 01 | 28.94% |
| | Business | 04 | 25 | 02 | 02 | 01 | 02 | 30.55% |
| | Entertainment | 04 | 04 | 24 | 03 | 02 | 00 | 35.13% |
| | Health | 08 | 03 | 01 | 27 | 02 | 01 | 35.71% |
| | Spiritual | 02 | 01 | 00 | 01 | 15 | 02 | 28.57% |
| | Astrology | 02 | 02 | 01 | 03 | 02 | 26 | 27.77% |
| **Average error rate for single experiment using 8-fold cross validation** | | | | | | | | 31.11% |

**Table 4.** 6-fold cross validation

| NB classifier predicted | | | | | | | | Error rate |
|---|---|---|---|---|---|---|---|---|
| Actual Category belongs | Category | Sports | Business | Entertainment | Health | Spiritual | Astrology | |
| | Sports | 28 | 2 | 3 | 4 | 2 | 1 | 30.00% |
| | Business | 9 | 29 | 3 | 2 | 2 | 4 | 40.81% |
| | Entertainment | 10 | 3 | 31 | 2 | 1 | 3 | 38.00% |
| | Health | 7 | 4 | 3 | 34 | 3 | 2 | 35.84% |
| | Spiritual | 2 | 3 | 1 | 3 | 33 | 5 | 29.79% |
| | Astrology | 0 | 4 | 2 | 2 | 4 | 29 | 29.26% |
| **Average error rate for single experiment using 6-fold cross validation** | | | | | | | | 33.95% |

**Table 5.** 4-fold cross validation

| NB classifier predicted | | | | | | | | Error rate |
|---|---|---|---|---|---|---|---|---|
| Actual Category belongs | Category | Sports | Business | Entertainment | Health | Spiritual | Astrology | |
| | Sports | 49 | 11 | 5 | 4 | 2 | 3 | 33.78% |
| | Business | 9 | 46 | 3 | 6 | 3 | 4 | 35.21% |
| | Entertainment | 13 | 5 | 39 | 3 | 1 | 2 | 38.09% |
| | Health | 5 | 9 | 3 | 51 | 7 | 6 | 37.03% |
| | Spiritual | 3 | 9 | 6 | 1 | 46 | 3 | 32.35% |
| | Astrology | 1 | 7 | 2 | 7 | 9 | 37 | 41.26% |
| **Average error rate for single experiment using 4-fold cross validation** | | | | | | | | 36.28% |

**Table 6.** 2-fold cross validation

| NB classifier predicted | | | | | | | | Error rate |
|---|---|---|---|---|---|---|---|---|
| Actual Category belongs | Category | Sports | Business | Entertainment | Health | Spiritual | Astrology | |
| | Sports | 105 | 13 | 26 | 7 | 5 | 8 | 35.97% |
| | Business | 20 | 82 | 17 | 9 | 3 | 3 | 38.80% |
| | Entertainment | 27 | 19 | 86 | 5 | 7 | 3 | 41.49% |
| | Health | 11 | 18 | 9 | 99 | 10 | 14 | 38.50% |
| | Spiritual | 7 | 16 | 10 | 13 | 67 | 8 | 44.62% |
| | Astrology | 5 | 13 | 6 | 15 | 12 | 62 | 45.13% |
| **Average error rate for single experiment using 2-fold cross validation** | | | | | | | | 40.75% |

Based on experiment of nb using feature selection, we have obtained the average accuracy of 88.96% from k-fold for 60 numbers of terms. based on our result analysis of table 7, we conclude that by selecting limited number of influence terms for particular category increase the performance of classifier.

## 5. Applications

- Spam categorization or spam filtering: classify email from its general category to identify fraud or tempter emails.
- Email routing: This is an automation process to send an incoming email to a specific target email address based on its subject or content written in body (topic dependent).
- Language detection, identification and content separation from multilingual documents.
- Genre categorization: Automatic identification of different genre from the written document or text.
- Readability evaluation: Measure the degree of readability based on the complexity of vocabulary, font style and size, line spacing, various grammatical form, reading speed etc.

**Table 7.** Error rate of NB using features selection

| Sr. No. | k-fold / # of terms | 10-fold | 8-fold | 6-fold | 4-fold | 2-fold | Average Error rate |
|---|---|---|---|---|---|---|---|
| 1 | 20 | 86.50 | 86.01 | 85.90 | 83.01 | 81.04 | 15.51% |
| 2 | 40 | 90.56 | 90.40 | 90.42 | 88.49 | 84.90 | 11.05% |
| **3** | **60** | **90.80** | **90.39** | **90.29** | **88.20** | **85.11** | **11.04%** |
| 4 | 80 | 88.34 | 87.98 | 87.89 | 86.00 | 82.30 | 13.50% |
| 5 | 100 | 86.08 | 86.76 | 86.70 | 84.45 | 82.00 | 14.80% |
| 6 | 120 | 85.89 | 85.03 | 84.70 | 81.08 | 79.00 | 16.86% |
| 7 | 140 | 84.90 | 84.00 | 83.67 | 81.90 | 78.70 | 17.37% |
| 8 | 160 | 82.00 | 82.11 | 82.02 | 79.70 | 76.90 | 19.45% |
| 9 | 180 | 82.01 | 81.00 | 80.45 | 78.00 | 75.01 | 20.71% |
| 10 | 200 | 81.22 | 80.99 | 80.00 | 78.05 | 76.90 | 20.57% |
| 11 | 300 | 80.02 | 79.70 | 79.02 | 76.90 | 74.00 | 22.07% |
| 12 | 400 | 79.23 | 79.00 | 78.34 | 76.10 | 74.20 | 22.63% |
| 13 | 500 | 78.00 | 78.01 | 77.86 | 76.03 | 72.90 | 23.44% |
| 14 | 600 | 77.68 | 77.00 | 77.02 | 74.90 | 71.07 | 24.47% |
| 15 | 700 | 77.00 | 75.89 | 75.09 | 73.00 | 70.01 | 25.80% |
| 16 | 800 | 75.06 | 73.30 | 73.00 | 70.99 | 69.06 | 27.72% |
| 17 | 1000 | 72.23 | 70.90 | 70.45 | 68.07 | 66.09 | 30.45% |

- Document indexing.
- Categorization can also be important for other applications like, survey coding, document clustering, authorship attribution etc.

## 6. Conclusion and Future Work

This work has been carried out to Gujarati document classification using Naïve Bayes classifier. We have also discussed the results of classifier for multi-category Gujarati documents. NB classifier performance is checked based on feature selection TF-IDF and without feature selection technique. We have used k-fold cross validation to evaluate the performance of NB classifier. In this research we have considered value of k = 2, 4, 6, 8 and 10. We obtained minimum error rate in 10-fold cross validation and maximum error rate in 2-fold cross validation. We achieved maximum accuracy of 88.96% and 75.74% using feature selection and without using feature selection technique respectively. We achieved good accuracy by considering 60 terms in TF-IDF which is more influence and related to the particular domain specific category. NB classifier consider each word as an independent word in document and needs training to implement. In future we will apply Ontology based approach for Gujarati document classification and extend this work by adding new category in Ontology which can be used in other research in area of Natural Language Processing and Mining.

## 7. References


1. Lin SH, Chen M C, Ho JM, Huang YM. ACIRD: Intelligent Internet document organization and retrieval. IEEE Transactions on Knowledge and Data Engineering. 2002; 14(3):599–614. https://doi.org/10.1109/TKDE.2002.1000345
2. Lee LH, Isa D. Automatically computed document dependent weighting factor facility for Naïve Bayes classification. Expert Systems with Applications, 2010; 37(12):8471–8. https://doi.org/10.1016/j.eswa.2010.05.030
3. Zhang H. The Optimality of Naive Bayes. Barr V, Markov Z, editors. FLAIRS Conference; AAAI Press; 2004.
4. Patil JJ, Bogiri N. Automatic text categorization Marathi documents. International Journal of Advance Research in Computer Science and Management Studies. 2015; 3(3):280–7. https://doi.org/10.1109/icesa.2015.7503438
5. Patil M, Game P. Comparison of Marathi text classifiers. ACEEE International Journal on Information Technology. 2014; 4(1):11–22.

6. mandal ak, sen r. supervised learning method for bangla web Document Categorization. International Journal of Artificial Intelligence and Applications. 2014; 5(5):93–105. https://doi.org/10.5121/ijaia.2014.5508
7. Murthy VG, Vardhan BV, Sarangam K, Reddy PVP. A comparative study on term weighting methods for automated Telugu text categorization with effective classifiers. International Journal of Data Mining and Knowledge Management Process. 2013; 3(6):95. https://doi.org/10.5121/ijdkp.2013.3606
8. Swamy MN, Hanumanthappa M. Indian language text representation and categorization using supervised learning algorithm. International Journal of Data Mining Techniques and Applications. 2013; 2:251–7.
9. Naseeb N, Gupta V. Domain based classification of punjabi text documents using ontology and hybrid based approach. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing COLING; 2012. p. 109–122.
10. Rajan K, Ramalingam V, Ganesan M, Palanivel S, Palaniappan B. Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Systems with Applications. 2009, 36(8):10914–8. https://doi.org/10.1016/j.eswa.2009.02.010
11. Raghuveer K, Murthy KN. Text categorization in Indian languages using machine learning approaches. IICAI; 2007. p. 1864–83.
12. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2002; 10:79–86.
13. Rogati M, Yang Y. High-performing feature selection for text classification. Proceedings of the 11th International Conference on Information and Knowledge Management; 2002. p. 659–61. https://doi.org/10.1145/584792.584911
14. Forman G. An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research. 2003; 3:1289–305.
15. Tan S, Zhang J. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications. 2008; 34(4):2622–9. https://doi.org/10.1016/j.eswa.2007.05.028
16. Prabowo R, Thelwall M. Sentiment analysis: A combined approach. Journal of Informetrics. 2009; 3(2):143–57. https://doi.org/10.1016/j.joi.2009.01.003
17. Alsaleem S. Automated Arabic text categorization using SVM and NB. International Arab Journal of e-Technology. 2011; 2(2):124–8.
18. El Kourdi M, Bensaid A, Rachidi TE. Automatic Arabic document categorization based on the Naïve Bayes algorithm. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Association for Computational Linguistics; 2004. p. 51–8. https://doi.org/10.3115/1621804.1621819
19. Hadni M, Lachkar A, Ouatik SA. A new and efficient stemming technique for Arabic text categorization. 2012 International Conference on Multimedia Computing and Systems (ICMCS); 2012. p. 791–6. https://doi.org/10.1109/ICMCS.2012.6320308
20. Harrag F, El-Qawasmah E, Al-Salman AMS. Stemming as a feature reduction technique for Arabic text categorization. 2011 10th International Symposium on Programming and Systems (ISPS); 2011. p. 128–33.
21. Halder T, Karforma S, Mandal R. A novel data hiding approach by pixel-value-difference steganography and optimal adjustment to secure e-governance documents. Indian Journal of Science and Technology. 2015 Jul; 8(16):1–7. https://doi.org/10.17485/ijst/2015/v8i16/51269
22. Prakash KB. Mining issues in traditional Indian web documents. Indian Journal of Science and Technology. 2015 Nov; 8(32):1–11.
23. Antipov KV, Vinokur AI, Simakov SP, Isakov YV, Kazakova AY. Digitization of Russian parish registers of the 18-20th centuries as the contribution to the cultural foundation of historical documents. Indian Journal of Science and Technology. 2015 Dec; 8(10):1–10. https://doi.org/10.17485/ijst/2015/v8is(10)/87462
24. Posonia AM, Jyothi VL. Context-based classification of XML documents in feature clustering. Indian Journal of Science and Technology. 2014 Jan; 7(9):1–4.
25. Karthika S, Sairam N. A naïve bayesian classifier for educational qualification. Indian Journal of Science and Technology. 2015 Jul; 8(16):1–5. https://doi.org/10.17485/ijst/2015/v8i16/62055
26. Sarangi PK, Ahmed P, Ravulakollu KK. Naïve Bayes classifier with LU factorization for recognition of handwritten Odia numerals. Indian Journal of Science and Technology. 2014 Jan; 7(1):1–4.