I. J. Mathematical Sciences and Computing, **2023, 1, 13-22**
Published Online on February 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2023.01.02

Comparison of Machine Learning Algorithms in Domain Specific Information Extraction

M. Rajasekar*
Hindustan Institute of Technology and Science, Chennai, India
Email: sekarca07@gmail.com
https://orcid.org/0000-0003-1084-7881
*Corresponding Author

Angelina Geetha
Hindustan Institute of Technology and Science, Chennai, India
Email: angelinag@hindustanuniv.ac.in

Received: 28 October, 2021; Revised: 24 November, 2021; Accepted: 20 December, 2021; Published: 08 February, 2023

Abstract: Information Extraction is an essential task in Natural Language Processing. It is the process of extracting useful information from unstructured text. Information extraction helps in most of the NLP applications like sentiment analysis, named entity recognition, medical data extraction, features extraction from research articles, feature extraction from agriculture, etc. Most of the applications in information extraction are performed by machine learning models. Many research work shave been carried out on machine learning based information extraction from various domain texts in English such as Bio medical, Share market, Weather, Business, Social media, Agriculture, Engineering, and Tourism. However domain specific information extraction for a particular regional language is still a challenge. There are different types of classification algorithms. However, for a selected domain to select the appropriate classification algorithm is very difficult. In this paper three famous classification algorithms are selected to do information extraction by classifying the Gynecological domain data in Tamil Language. The main objective or this research work is to analyze the machine learning methods which is suitable for Tamil domain specific text documents. There are 1635 documents being involved in classification task to extract the features by these selected three algorithms. By evaluating the classification task of each model it has been found that the Naive Bayes classification model provides highest accuracy value (84%) for the gynecological domain data. The F1-Score, Error rate and Execution time also evaluated for the selected machine learning models. The evaluation of performance has proved that the Na ïve Bayes classification model gives optimal results. It has been concluded that the Na ïve Bayes classification model is the best model to classify the gynaecological domain text in Tamil language

Index Terms: Machine Learning, Information Extraction, Gynecology, Naive Bayes classifier, Support Vector Machines, K-nearest neibhour classifier.

1. Introduction

Since the recent decade the text information in the form of digital way has been booming in our Technological world. As it is significant, it is needed to bring together and classify this text data for further development. The text information is to be extracted, classified, and clustered according to the categories. Mostly information is available in the global language, English. Information is available in unstructured or semi structured format. The information is to be processed to get in structured format. The main goal of text extraction is to enable users to extract information from textual resources and deals with operations like retrieval, classification (supervised, unsupervised, unsupervised and semi supervised) and summarization [1]. Several approaches have been done in information extraction from English text. But in India, a multi lingual country, information is scattered in many languages. It is a big challenge to extract the needed information in the regional languages. But there are a number of research contributions by various researchers in information extraction from a specific domain text in various regional languages. To classify the unstructured information is one of the tasks in information extraction. Based on the categories of a particular domain, information is to be classified. There are many machine learning based classification models that are available. To select the appropriate classification model for the selected domain is a challenging job. Text data is available in various domains like medical records, agriculture, social media, climate and weather, marketing, finance, legal documents, and research articles. To extract the useful information from domain data is the mandatory task in several artificial intelligence applications.

Disease predictions, disease symptom analysis, marketing predictions, weather prediction, agricultural prescriptions, climate analysis and many more applications are built by using extracted structured data. In recent years medical field has grown up with artificial intelligent development. In India gynecological diseases like breast cancer, vaginal cancer, cervical cancer, and ovarian cancer are increasing rapidly. The main reason for this is rural women have lack of adequate knowledge in their health issues. This ratio is increased in well educated women also. There are such research work has been done in this gynecological domain. Traditional remedies for women health issues are viewed using machine learning method is proposed in E Marutuvachi" – Information Extraction Framework for data about obstetrics and gynecology in Tamil [2]. In this research paper the gynecological data document in Tamil language is used to implement the information extraction to extract useful information for rural women in Tamilnadu. Based on the extracted information the user(rural women) can get adequate knowledge about their health issues. So they can be aware about the gynecological diseases to prevent them. This paper deals with the machine learning models to extract useful features from Gynecological domain text in Tamil language and compares the performance of those classification models.

2. Background Study

Information classification task is the major task in extracting useful information from unstructured data. The text classification will be performed by most of the classification models. There are various types of machine learning classification models that are available for text classification. These classification algorithms classify the input data into a specific category. The classification algorithms are classified as Binary classification, multi-class classification, and multi-label classification. Binary classification means that the classification task gives two possible outcomes. More than two possible outcomes are referred to as multi-class classification. In multi-label classification each sample is mapped to a set of target labels.

The following is a list of machine learning classification algorithms are Linear Regression, Naive Bayes classification, Stochastic Gradien Descent, K-Nearest Neighbours, Decision Tree, Random Forest, Support Vector Machine.

The performances of the above machine learning classification models are verified to choose the best model to extract the information by classification task.

2.1. Gynecological domain data in Tamil

All the above machine learning algorithms have been implemented in information classification task in English. In this research work, the gynecological domain data in Tamil language is chosen to process the extraction of useful information. 1635 documents in gynecological domain data in Tamil language have been collected from various internet sources web sites, blogs, News. These documents are categorized as women physical problems. The content of the document is available based on the category of the document.

3. Literature Review

An efficient English text classification using selected machine learning technique has been approached by Xiaou [3]. In this research work, the support vector machine model is used to classify English text and documents. Two analytic experiments have been done to check the selected classifiers to classify the English text. The Rocchio classifier provides the best performance results in 1033 English text documents compared with SVM classifier. The Rocchio classifier provides a good accuracy of 90% in English documents classification.

An Automatic text classification using machine learning and optimization algorithms has been proposed by R. Janani &S. Vijayarani [4]. The text classification task is done in two phases. The first one is to select important features

for classification and the second one is to classify the text documents. An optimization technique for feature selection algorithm is used to implement the classification of text documents. To evaluate the proposed method the performance of selected algorithm is compared with other classification techniques SVM classifier, K-nearest neighour, Naive Bayes algorithm, and probabilistic neural networks. The proposed method has yielded very high accuracy compared with the other classification models.

A comparative study of five text classification algorithms with their improvements has been done by Ahmed et al.,[5]. This research work has been done on the comparative study of all the amendments done on five essential text classification algorithms, namely Decision Tree, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, and Hidden Markov Model. The comparative study has been done on Learner modification, main algorithm modifications and addition in feature extraction and reduction.

A Performance comparison of different classification algorithms for household poverty classification has been done by Janelyn et al.,[6]. To identify the poor households of poverty alleviation programs they have used machine learning algorithms to analyze the poverty data from Community based monitoring system database of Langangilang, Abra, Philippines. Major classification algorithms ~~are compared~~ like Naive Bayes, ID3, Decision Tree, Logistic Regression and K-Nearest Neighbour have been compared. The algorithmshave been evaluated based on the performance metrics--accuracy, precision, recall, F1 score, error date, and AUC. This study concludes that the Naive Bayes algorithm is the most efficient algorithm for the prediction of households that are poor or non-poor. The error rate of the Naive Bayes algorithm is 0.014 only.

The Automated Named Entity Recognition model has been proposed by R. Srinivasan et al.,[7]. By using supervised learning method, the system automatically assigns the NER tags to the tokens in Tamil documents. The Naïve Bayes algorithm is used to design the feature extraction from Tamil documents. The corpus has 1028 documents in Tamil language. They have used three feature extraction models such as Regex Feature extraction, Morphological Feature extraction, and Context Feature extraction. The model has been evaluated by Precision and Recall method. The final F-measure has achieved 83.54%.

A framework for Named Entity Recognition for Malayalam has been proposed by R. Rajimol et al.,[8]. It is deep learning approach in NER system for Malayalam. A comparison has been done across the different deep learning methods, recurrent neural networks, gated recurrent unit, long short-term memory, and bi-directional long short-term memory. For NER, it has been discovered that DL-based approaches outperform traditional shallow-learning-based approaches significantly. It has been found that deep learning approach for NER system in Malayalam has outperformed in terms of precision, recall, and F-measure and it has yielded an F-score of8.92%.

A Rule based Kannada named entity recognition has been approached by M. Pushpalatha et al.,[9]. There are two stages--pre-processing and entity recognition. In the pre-processing stage the sentences are collected from different sources and are divided into words. For the named entity recognition of Kannada words, the Support Vector Machine model is used. The rule is formed to identify the names of the person, place, and designation of the person. This approach has yielded a good result of 89.32% in accuracy.

From the above review, it has been noted that there is an essential gap in the comparison of machine learning algorithms in Tamil language domain datasets in performance metrics like accuracy, precision, recall, F1-score, error rate and time duration. In this research work the above said performance metrics are compared with the essential classification algorithms, namely Naive Bayes classification algorithm, support vector machines, and K-nearest neighbour algorithm.

3.1. Objectives

- To analyze the performance of classification methods for Tamil domain text.
- The check the classification methods which is most suitable Tamil domain specific text.
- To compare the performance of machine learning classification algorithms for gynaecological domain text in Tamil Language.
- To prepare the Gynecoloical Tamil text corpus for text analysis.

The comparison of machine learning algorithms is done based on performance metrics, accuracy, precision, recall, F1-score, error rate, and time duration for classification.

4. Materials and Methods

The classification task is done by using gynaecological domain datasets in Tamil Language. The corpus is developed using blogs, news websites. Collected data is so noisy. So it is pre-processed using cleansing and cleaning techniques. The data is cleaned by removing punctuation marks, extra spaces, English words, and hyperlinks. The cleaned datasets are stored with their appropriate keywords to access easily. There are 9 types of categories of 1635 documents that have been collected for classification. The list of documents is shown in Table.1.

Table 1. List of documents

Topic	**No. of Doc**
ஆரோக்கியபாடம்(Women Health Education)	211
வெள்ளைப்படுதல்(White discharge)	167
சினைப்பைபுற்றுநோய்(Ovarian Cancer)	127
கருப்பைநீர்க்கட்டிகள்(PCOS/PCOD)	231
பிறப்புறுப்புபுற்றுநோய்(Vaginal Cancer)	195
கருப்பைவாய்புற்றுநோய்(Cervical Cancer)	168
மார்பகப்புற்றுநோய்(Breast Cancer)	206
தைராய்டு(Thyroid)	153
மாதவிடாய்பிரச்சினை(Menstrual Problems)	177
Total number of documents	**1635**

This collection of documents is used to check the accuracy of the classification done manually. By using machine learning classification techniques the classification is checked for the accuracy of the classified documents.

The steps to extract the useful information from the collected text are as follows:

- Text Pre-processing
- Morphological Analysis
- Part of Speech Tagging
- Named Entity Recognition
- Domain Specific named entity recognition
- Keywords generation using Named Entity tags
- Develop Information Extraction Model to Extract data
- Implement Machine learning models to extract data
- Comparison of Machine Learning algorithms

Using these tasks the information extraction process has been done by step by step. For these processes the materials are most important elements.

4.1. Machine learning based classification models

There are many text classification algorithms being available a few of which are most essential and famous [10]. They are Logistic Regression, Naive Bayes classifier, Stochastic Gradient Descent, K-Nearest Neighbour, Decision Tree, Random Forest, and Support Vector algorithm. These algorithms are briefly explained here.

4.1.1. Logistic Regression

In this classification algorithm, the probabilities of classified data describing the possible outcomes of a single modeled probability logistic function. It is mostly used for several independent variables and single outcome variable. The limitation of this model is that it can only predict the binary variable.

4.1.2. Naive Bayes classification

Naive Bayes classification algorithm is based on Naive Bayes theorem, the probabilistic assumption of individuality of every data instance. The Naive Bayes model performs well in document classification. To find the most probable classification tag of the given entity,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

is used.

Finding the probability of A, when probability of B is true, P(A|B) is Priori probability. P(B|A) is posterior probability.

4.1.3. Stochastic Gradient Descent

The gradient descent optimization algorithm will do iteration over a function and adjust the parameters of the function until it finds the minimum combinations. This SGD model is very easy to implement and a very efficient method to fit in linear models. It supports different loss functions and penalties for classification. This model requires a number of hyper-parameters and it is very sensitive to feature scaling.

4.1.4. K-Nearest Neighbour algorithm

The K-NN algorithm is comparability based learning algorithm and it has been a very effective model for text categorization. Given a text document, the K-NN algorithm finds the K nearest neighbour element among the whole training documents. It uses the categories of K-NN to calculate the weight of neighbour candidate. The similarity score of each document with the selected word is to be calculated and the minimum value of neighbour weight is to be found. The resulting weighted sum is used as neighbour score. It is to be obtained for the test document. Based on the weighted value the word is to be categorized.

4.1.5. Decision Tree

Given data is classified by applying sequence of rules that can be implemented till the model produces optimum results. The decision tree classification uses sequence of conditions with two possible outcomes. Each outcome may have another condition to classify the data. It is simple to understand and easy to visualize.

4.1.6. Random Forest

Random Forest model is a moderate predictor, established that to fits a number of decision trees on various datasets. It uses the average value of prediction to improve the accuracy of the model. It also controls the over-fitting data to improve the accuracy. The sub-datasets are also the same as actual input datasets. It is more accurate than decision trees in the classification task. It is difficult to implement because it has complex algorithm process.

4.1.7. Support Vector Machine

Support Vector Machine or SVM is, perhaps, the most famous Supervised Learning model, which is utilized for Classification and Regression issues. Essentially, it is used for Classification issues in Machine Learning. The objective of the SVM algorithm is to make the best line or decision limit that can isolate n-dimensional space into classes so we can easily put the new data in the right classification later on. This best choice limit is known as a hyper plane. SVM model is used for text classification, image classification, face detection, and identifying objects. There are two types of SVM classifier-- Linear and Non-Linear model. If data is to be classified as two separate classes by single linear plots, it is called linear. If it is not possible to separate data as two classes, it is non-linear model. In SVM model the data to be considered as word vectorization, it is to convert text documents into numerical feature vectors.

4.2. TF-IDF weighting method

To classify a document in a particular category the TF-IDF method is used tofind the frequency of the word in the document. Based on the weight value of TF-IDF, the document is classified as the particular category. The most popular method to convert word to vector is TF-IDF (term frequency-inverse document frequency). TF refers to the frequency of occurrence of a selected word in the document. IDF refers to down scales words appear in a lot of documents.

$$tf - idf(t,d) = tf(t,d) * id(t) \tag{2}$$

$$idf(t) = log\left\{\frac{n}{df(t)}\right\} + 1, (if\ smoot.\_idf\ = false) \tag{3}$$

5. Implementation and Evaluation

The information extraction implementation steps are,

- Morphological Analysis
- Part of Speech Tagging
- Named Entity Extraction
- Relation and Keyword Extraction
- Information Extraction Model Development
- Comparison of Machine Learning Models for IE Task

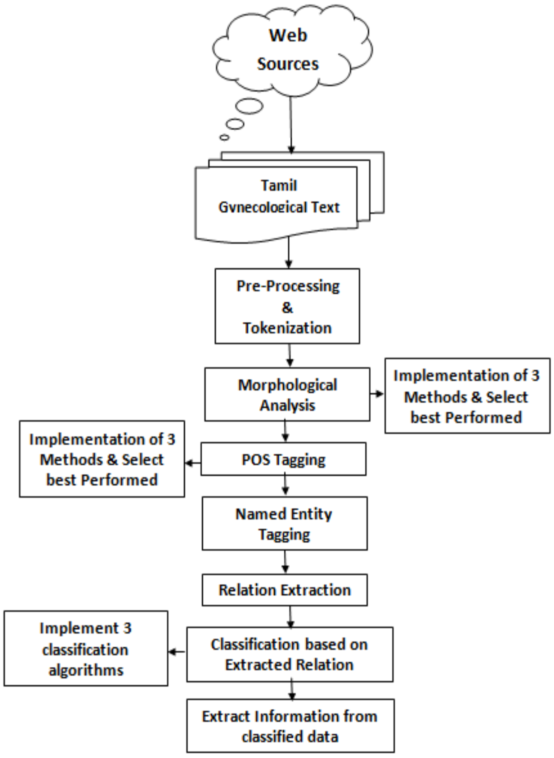

Web Sources
Tamil Gynecological Text
Pre-Processing & Tokenization
Morphological Analysis
Implementation of 3 Methods & Select best Performed
POS Tagging
Implementation of 3 Methods & Select best Performed
Named Entity Tagging
Relation Extraction
Classification based on Extracted Relation
Implement 3 classification algorithms
Extract Information from classified data

Fig. 1. Overall IE Model Design

The abrove figure explains the process of Information Extraction.

The collected corpus is used to extract the useful information. The first process is to clean the data using preprocessing, such are removing unwanted characters, special characters, and hyperlinks. After cleaning the text corpus it is tokenized as word by word. Then it is analyzed morphologically using machine learning algorithms. Then the corpus is tagged with part of speech tags using POS tagging models for the selected domain.

From the tagged corpus the named entities are collected to extract the keywords to search from the corpus. Using the Keywords the relations are created for all the documents in the corpus. For these three machines learning algorithms are used to extract the useful information in a structured format.

The corpus data is implemented in the classification task in three main classification algorithms -- Naive Bayes classification, Support vector machine, and K-nearest neighbour classification. The classification algorithms are compared by the performance metrics, namely accuracy, precision, recall, F1-Score, error-rate, and time duration.

To check the accuracy of the classification model the following formula is used:

$$Accuracy = \frac{n}{N}$$

where,
n – number of correctly predicted documents
N – number of documents in the corpus

To find out the precision, recall, and F1-Score values the following formula is used. The precision and recall evaluation method is based on the following confusion matrix:

Table 2. Confusion Matrix

	Negative (Predicted)	Positive (Predicted)
Negative (Actual)	True negative	False positive
Positive (Actual)	False negative	True positive

Based on the confusion matrix the precision and recall values are calculated by the formula

$$Precision = \frac{truepositives}{truepositives + falsepositives} \tag{4}$$

$$Recall = \frac{truepositives}{truepositives + falsenegatives} \tag{5}$$

And, finally the F-Score is calculated as follows

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

Error rate is equal to false negative value in the confusion matrix.

And finally, run time duration of the classification task is calculated by using the following code:

$$execution\_time = timeit.timeit(code, number = 1)$$

The evaluation values are calculated as follows.

5.1 Accuracy

Accuracy values of the three classification algorithms are shown in Table.3.

Table 3. Accuracy of classification algorithms

Topic	**Accuracy**		
	Na ïve Bayes	**KNN**	**SVM**
ஆரோக்கியபாடம்(Women Health Education)	86.25592417	81.04265403	84.83412
வெள்ளைப்படுதல் (White discharge)	84.43113772	83.83233533	82.63473
சினைப்பைபுற்றுநோய்(Ovarian Cancer)	72.44094488	79.52755906	74.80315
கருப்பைநீர்க்கட்டிகள்(PCOS/PCOD)	84.84848485	83.11688312	81.81818
பிறப்புறுப்புபுற்றுநோய் (Vaginal Cancer)	78.97435897	73.33333333	73.84615
கருப்பைவாய்புற்றுநோய்(Cervical Cancer)	82.73809524	84.52380952	88.09524
மார்பகப்புற்றுநோய்(Breast Cancer)	81.55339806	73.30097087	77.6699
தைராய்டு(Thyroid)	79.73856209	84.31372549	79.08497
மாதவிடாய்பிரச்சினை(Menstrual Problems)	89.26553672	74.57627119	82.48588
Overall Accuracy	**82.6911315**	**79.57186544**	**80.73394**

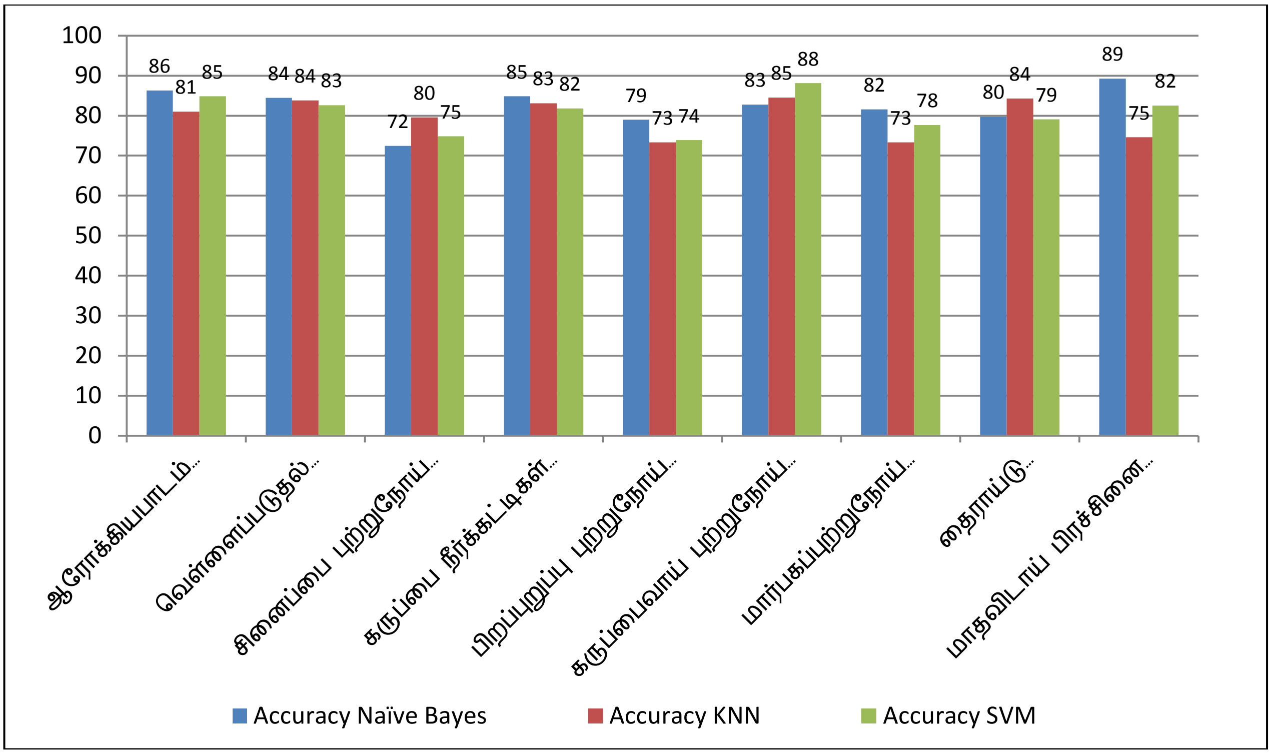


Fig. 2. Accuracy comparison chart of classification algorithms

5.2. *Precision, Recall, F1-Score*

The precision, recall, and f1-score comparison of classification algorithms is shown in Table.4.

Table 4. Precision, Recall, and F1-Score of Classification algorithms

text_cat_id	**No_docs**	**Na ïve Bayes**			**KNN**			**SVM**		
		Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
ஆரோக்கியபாடம் (Women Health Education)	211	0.93	0.81	0.87	0.87	0.75	0.81	0.90	0.78	0.83
வெள்ளைப்படுதல் (White discharge)	167	0.96	0.77	0.86	0.90	0.71	0.80	0.93	0.74	0.82
சினைப்பை புற்றுநோய் (Ovarian Cancer)	127	0.94	0.79	0.80	0.88	0.73	0.74	0.91	0.75	0.77
கருப்பை நீர்க்கட்டிகள் (PCOS/PCOD)	231	0.99	0.87	0.93	0.93	0.81	0.87	0.96	0.83	0.89
பிறப்புறுப்பு புற்றுநோய் (Vaginal Cancer)	195	0.97	0.84	0.90	0.91	0.78	0.84	0.94	0.80	0.87
கருப்பைவாய் புற்றுநோய் (Cervical Cancer)	168	0.97	0.82	0.89	0.91	0.76	0.83	0.93	0.78	0.85
மார்பகப்புற்றுநோய் (Breast Cancer)	206	0.99	0.83	0.90	0.93	0.77	0.84	0.96	0.79	0.87
தைராய்டு (Thyroid)	153	0.98	0.76	0.86	0.92	0.70	0.80	0.95	0.72	0.82
மாதவிடாய் பிரச்சினை (Mentrual Problems)	177	0.99	0.84	0.91	0.93	0.78	0.85	0.95	0.81	0.87
Overall Values	**1635**	**0.97**	**0.81**	**0.88**	**0.91**	**0.75**	**0.82**	**0.94**	**0.78**	**0.85**

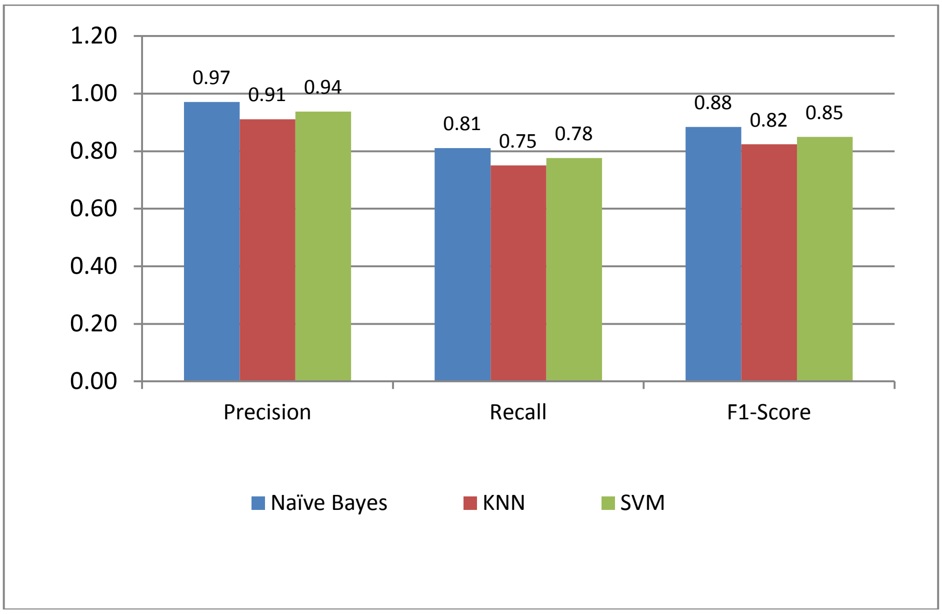


Fig. 3. Precision, Recall, and F1-Score chart of classification algorithms

5.3. *Error rate*

The error rate of each classification model is evaluated by its confusion matrix value, false negative. Error rates of all the three models are shown in Table 5.

Table 5. Error rate

Model	no.docs	FN	Error_rate
Na ïve Bayes	1635	21	1.28440
KNN	1635	63	3.85321
SVM	1635	47	2.87462

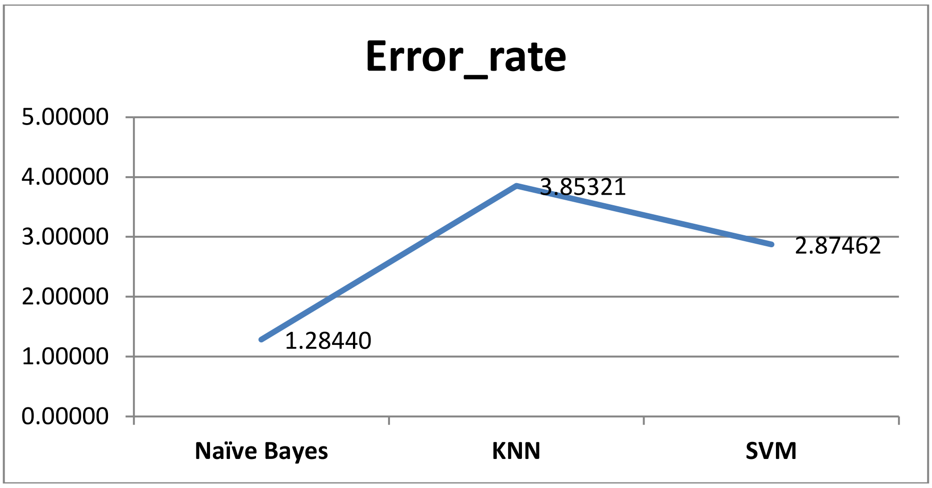


Fig. 4. Error rate chart

5.4. *Time duration*

The time duration to classify the given dataset by classification algorithm is evaluated in the beginning stage. With 10 documents each algorithm is tested for time duration to execute the code block. The time duration taken in Naive Bayes model is 0:00:00:0014350. For KNN model it is 0:00:00:0018945 and for SVM model it is 0:00:00:0017453. Only the Naive Bayes model has taken the lowest time when compared with the other two classification models.

5.5 *Overall performance comparison*

The performances of all the three classification algorithms are compared with metrics, accuracy, precision, recall, f1-score, error rate, and time duration. By comparing the performances of all the metrics, the Naive Bayes model has been proved to be the most suitable model to perform the classification task in gynaecological domain data in Tamil language. The results of the comparison of all the performances shown in Table.6.

Table 6. Overall performance comparison

Model	Accuracy	F1-Score	Error_rate	exec_time
Na ïve Bayes	82.69113	0.88348	1.28440	0:00:00:0014350
KNN	79.57187	0.82348	3.85321	0:00:00:0018945
SVM	80.73394	0.84948	2.87462	0:00:00:0017453

6. Conclusion

The text classification is an essential task in any Natural language processing. There are many applications based on classification such as Sentiment analysis, data prediction, and Feature extraction. Text classification has been done by using many famous classification algorithms for the global language, English. In this research work a comparison of machine learning based classification algorithms has been done to classify the gynaecological domain data in Tamil language. The selected algorithms are Na ïve Bayes classifier, k-nearest neighbor classifier, and support vector machines. The performance metrics of these classification algorithms are accuracy, precision, recall, f1-score, error rate, and time duration. The performances have been assessed for three classification algorithms based on gynaecological domain data in Tamil language. Based on the evaluation the Na ïve Bayes model have achieved high accuracy (82.6%). Based on the general machine learning methods evaluation the Na ïve Bayes model achieved high values in Precision, Recall and F1-Score(0.88). Then the error rate is very low for Na ïve Bayes model. Execution time also evaluated for the three models. The evaluation of performance has proved that the Na ïve Bayes classification model gives optimal results. It has been concluded that the Na ïve Bayes classification model is the best model to classify the gynaecological domain text in Tamil language.

References

[1] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar, "A Review Paper on Algorithms used for Tect Classification", International Journal of Application or Innovation in Engineering & Management (IJAIEM),ISSN 2319 - 4847, Volume 2, Issue 3, March 2013

[2] M. Rajasekar and Dr. A. Udhayakumar,"E Marutuvachi – Information Extraction Framework for data about obstetrics and gynecology in Tamil", Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018), IEEE Xplore, 2018

[3] Xiaoyu Luo, “Efficient English text classification using selected machine learning techniques”, Alexandria Engineering Journal, Volume 60, Issue 3, , Pages 3401-3409, 2021

[4] R. Janani & S. Vijayarani, “Automatic text classification using machine learning and optimization algorithms”, *Soft Computing* volume 25, pages1129–1145, 2021

[5] Ahmed H. Aliwy, Esraa H. Abdul Ameer, "Comparative Study of Five Text Classification Algorithms with their Improvements", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 14, pp. 4309-4319, 2017

[6] Janelyn A and Talingdan, "Performance comparison of different classification algorithms for household poverty classification", The proceedings of 2019 4th International Conference on Information Systems Engineering, IEEE Xplore, pp.11-15, doi:10.1109/ICISE, 2019

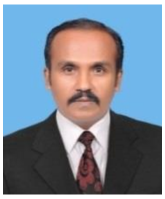
[7] R. Srinivasan and C.N. Subalalitha, “Automated Named Entity Recognition from Tamil Documents”, IEEE Proceedings of 1st International Conference on Energy, Systems and Information Processing (ICESIP), pp.1-5, 2019

[8] Rajimol and V.S. Anoop, “A Framework for Named Entity Recognition for Malayalam- A Comparison of different deep learning architectures”, Natural Language Processing Research, Vol.1(1-2) pp.14-22, 2020

[9] Pushpalatha and Dr. Anton Selvadoss Thanamani, “Rule Based Kannada Named Entity Recognition”, Journal of Critical Reviews, Vol. 7, Issue 4. 2020

[10] Ahmed H. Aliwy and Esraa H. Abdul Ameer, “Comparative Study of Five Text classification algorithms with their improvements”, International Journal of Applied Engineering Research ISSN 0973-4562, Vol.12, No.14, pp.4309-4319, 2017

Authors’ Profiles



M. Rajasekar is working as a Assistant Professor in the Department of Computer Applications in Hindustan Institute of Technology and Science, Padur, Chennai, India. He has received his Master in Computer Applications from Anna University, Chennai Tamilnadu in 2008. He has submitted his Ph.D. Thesis Repot in Hindustan Institute of Technology and Science, Chennai, Tamilnadu. His research interests are Natural Language Processing, Machine Learning Methods. Email: sekarca07@gmail.com

Dr. Angelina Geetha is working as a Professor and Dean – Engineering and Technology in Hindustan Institute of Technology and Science, Padur, Chennai, India. She has received her PhD from Anna University, Chennai, Tamilnadu in 2008. Her research interests are Machine Learning methods, Data mining, Information Retrieval.

How to cite this paper: M. Rajasekar, Angelina Geetha, "Comparison of Machine Learning Algorithms in Domain Specific Information Extraction", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.9, No.1, pp. 13-22, 2023. DOI: 10.5815/ijmsc.2023.01.02