Available online at http://scik.org

J. Math. Comput. Sci. 11 (2021), No. 6, 7140-7153

https://doi.org/10.28919/jmcs/6413

ISSN: 1927-5307

MACHINE LEARNING ALGORITHM FOR INFORMATION EXTRACTION FROM GYNAECOLOGICAL DOMAIN IN TAMIL

M. RAJASEKAR*, ANGELINA GEETHA

Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, India - 603 103

Copyright © 2021 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Information Extraction is a significant task in Natural Language Processing. It is the process of extracting useful information from unstructured text. Information extraction helps in most of the recent NLP applications like scientific research, financial investigation, business intelligence, media monitoring, healthcare records management, agriculture, and pharmacy research. There are several information extraction research approaches using many techniques from English dataset. As a multi lingual country India, it is actually challenging task to extract information from text in Indian language. Such research work has been done in the following domain data travel, food, agriculture, weather casting, social media, marketing and bio- medical. In this research work the relevant information extracted from Gynaecology related text data in Tamil language. The combination of machine learning based classification model and ontology representation is used extract the useful information. Auto filling IE framework is designed to extract the appropriate information in a structured format. The user query will be pre-processed and converted into entities to check from the classified data using ontological representation by using machine learning based classification model naïve bayes classification. From the ontological representation entity based relation extraction will be performed to fill the IE framework. The proposed IE framework given good results in extracting relevant information based on user query. It was analyzed for more than 57 user queries regarding gynaecological issues. The 75% of accuracy obtained for the correctness in user queries.

Keywords: machine learning; naive bayes classifier; ontology; information extraction; gynecology; entity relations.

2010 AMS Subject Classification: 68W40.

*Corresponding author

E-mail address: sekarca07@gmail.com

Received June 26, 2021

1. INTRODUCTION

In our current artificial world, the real world applications are evolving by ultimate master piece, information. Information is available in everywhere as many forms, unstructured, semi-structured and well structured. Mostly information is available in global language English. The information is to be available in well structured form to enable the users to use the data for their real time applications. There are several number of approaches have done in information extraction from English text. In India, the multi-lingual country the information scattered in many regional languages. The information extraction task in regional language is quite difficult. But the researchers have given their contribution in the area of extracting information from specific domain text in various regional languages. The IE research work done in various domain data, such as Tourism, Weather, Agriculture, Bio-medical, Marketing, Finance, Social media and agriculture. This paper illustrates the proposed unique approach in extracting relevant information from Gynaecological data in Tamil language.

This paper organized as Section 2: The literature review of related research works. Section 3: Brief study of machine learning, ontology and gynaecology. Section 4: Describes overall design of the proposed IE framework, Section 5: Describes evaluation and discussion and finally Section 6: Concluding the research work

2. BACKGROUND STUDY

The background study has been done in machine learning for classification, concepts of ontology for data representation, and Gynaecological statistics for domain data. Machine learning methods are used to classify the data items into category. The powerful data representation concept ontology is used to represent the data items as visually good.

2.1. Machine Learning models

Machine learning is an area of study of algorithms that learn from existing examples. To make a machine to handle data to learn themselves from existing examples. Machine learning methods are used to solve various problems like classifications, predictions, translate, analyzing and generating new models. The classification models in machine learning are logistic regression, artificial neural networks, random forest, naïve bayes, and K-nearest neighbor algorithm. In this paper, the classification task is done using naïve bayes model. From various machine learning classification algorithms, the naïve bayes classification model is selected to do the information extraction task with more accurate. Though there are many classification models available in machine learning approach, why the naive bayes classification approach is selected? The reasons are:

- Easy to train – required less training data
- Simplicity – This approach is more transparent, easy to apply
- Less memory needed
- It performs well even in more classification attributes

To find the most probable classification tag of the given entity,

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

Finding the probability of A, when probability of B is true, P(A|B) is Priori probability. P(B|A) is posterior probability. In such case if the probability of A is 0, means the classification task will be infinity. So, the Laplace smoothing will be involved. That is,

$$\theta_i = \frac{x_i + \alpha}{N + ad}$$

In simple language,

$$P(word) = \frac{word\ count + 1}{Total\ no.of\ words + No.of\ unique\ words}$$

Using naïve bayes classification model, the query dataset is classified based on the training phase classification. In training phase the classified datasets are involved in classification task to train the naïve bayes model.

2.2. Ontology

The concept of ontology is the study of categories of living or non-living things that exist in a particular domain. Ontology is an explicit specification of a conceptualization [1]. The general concept of ontology to represent the data as entities are attributes which relate to other entities using graphical representation. The smallest element in ontology is taxonomy. Taxonomy is the combination of vocabulary and structure of the domain language. Ontology is collection of rules, constraints and relationships with taxonomy. With the ontology, the instance is created by users. Here the instance is called user queries. By injecting instance into ontology the knowledge base (IE framework) is created.

In this research work, the ontology structure is created from human body. Women health issues were identified in a particular body part, it will relate with body part. The body part relates with body, like-wise the structure of ontology is developed. For example the women health issues were derived from the following ontological structure.

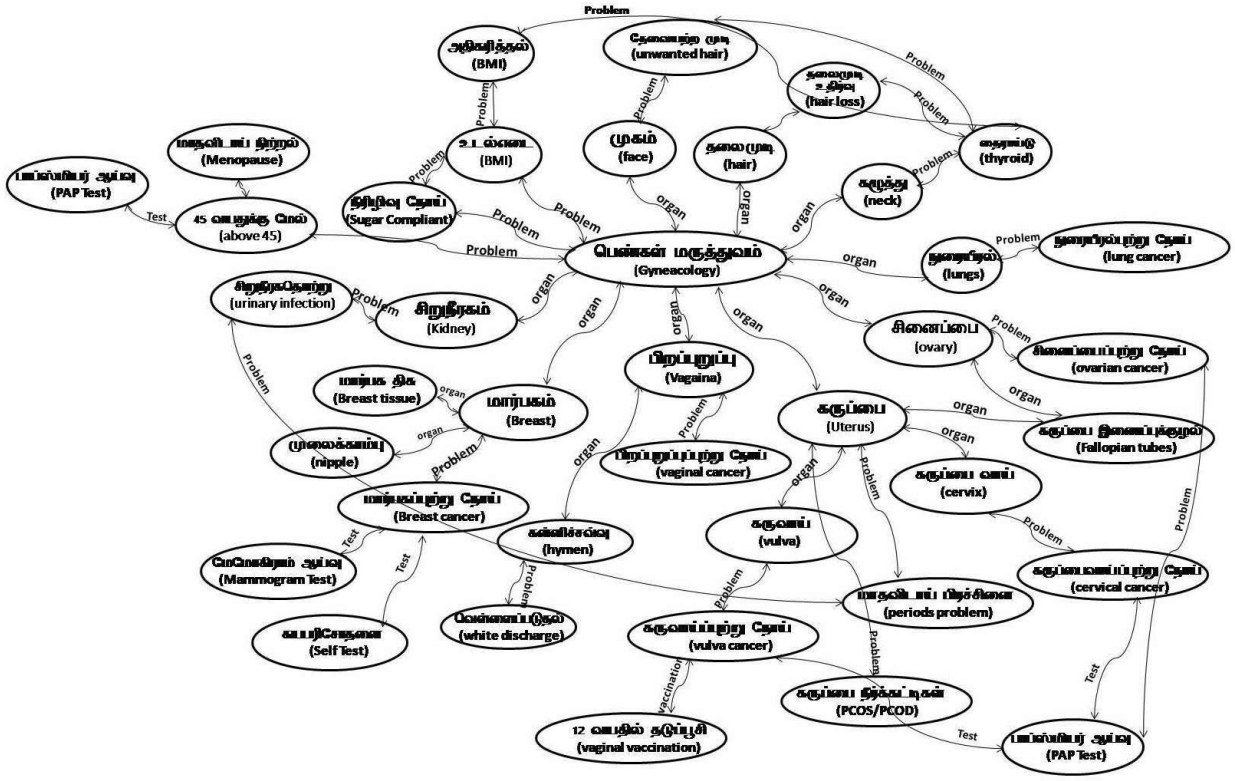

Figure.1 Ontological structure of Gynaecology

2.3. Gynaecology

In India, compared with men, women are more suffering from health issues. Particularly in rural area women have lack of knowledge about their health. According to the Cancer statistics of India 2020[2], there are 13, 92,179 cancer patients in India. Common types of cancers are breast, lung, mouth, cervix, uteri, and tongue. Most common type of issues in Indian women is breast cancer, cervical cancer, uterus cancer. Women in rural India don't have awareness of their health problems. The uterus cancer biggest issue will start from a tiny health issue, white discharge, urinal infections, and fungus infection. But they don't care about these small problems. Because they don't have adequate knowledge about it. They need a platform to share their health problems, to ask doubts and queries related to their health issues and get an idea about it.

All the information is available in Internet. They are scattered and mixed with unwanted data. This research work will give a platform to extract the useful and needed information from Tamil Gynaecological data. The user can ask queries about their health issues; the system will generate an Ontology tree by using the given keywords and give detailed information of the keywords.

3. RELATED WORKS

The reviews of existing works were done in Information extraction using ontology and machine learning methods in various domains.

An ontological framework for information extraction from academic data is proposed by Veena Jose *et al*[3]. This approach helps the academic search engine to perform effectively. The dynamically evolving ontology framework is implemented to help the academic search engine. The Word2Vec model is used to identify new keywords for development of ontology.

Machine learning based information extraction on tourism domain is proposed by Chantana Chantrapornchai *et al*[4]. The text classification and named entity recognition (NER) methodologies were used in Information extraction from tourism datasets. The machine learning tools such as SPACY and BERT were used to extract the specified domain data. It is found that the BERT model has given the best accuracy level (99%) for Named Entity Recognition. The BERT and SPACY models has given the accuracy around 95%-98% for Text classification task.

Analysis of Tweeter and RSS news feeds to perform sentiment analysis done by Shri Bharathi et al.[5]. The analysis was done in tweeter and RSS news fees comments and tweets. Two types of hypothesis used in this analysis. H0: Stock level indicator predicts the stock exchange values as 80% and above. H1: Stock level indicator along with Tweeter and RSS news feeds predicts the stock exchange value as accurately. The proposed approach improves the prediction values as 20% in prediction accuracy.

An algorithm for ontology based information extraction is designed by Peng Zhou *et al.*[6]. In the construction domain the energy requirements are extracted from energy conservation codes using the proposed algorithm. The combination of ontology-based pattern-matching extraction, text classification, domain-specific preprocessing, sequential dependency-based extraction, and cascaded extraction methods is used to solve the problems to extract the energy requirements. The algorithm is tested in extracting energy requirements from 2012 International Energy Conservation Codes. They have evaluated the method by Precision and Recall method. It provided the result as 97.4% recall and 98.5% precision.

The ontology based clinical data extraction is proposed by S Jusoh *et al.*[7]. The Ontology based clinical data extraction system is able to store, tacit and explicit the clinical data related to pediatric diseases as well as to supports clinicians for diagnosing and making decisions related to their patients. There are three ontology approaches implemented in the proposed model namely, source ontology, target ontology and mapping ontology. The source ontology is used to extract the clinical concepts from clinical

records. The target ontology is to represent the loading process. The mapping ontology makes relationship with source and destination ontology and converts the data into visualization of clinical data as template.

Ontology based Tamil-English cross lingual Information retrieval system proposed by D. Thenmozhi *et al*.[8]. The multi-lingual ontology constructed manually for the CLIR system. The main objective of the CLIR system using ontology is to remove ambiguity in Tamil query. The Tamil queries are transliterated into English and search in ontology representation for the needed information to be extracted. The agricultural text domain is selected to retrieve the information. This approach outperforms while compare to other methods in terms of precision.

A sentiment analysis of stock market prediction is done by Shri Bharathi et al[9]. It is the analysis process of Sensex points and really simple syndication news feeds for effective prediction. The domain data were collected from social media and stock market news. The domain stock market data is from ARBK Amman stock exchange. They have analysed the data without sentiment and with sentiment the moving average values. It is proved that with sentiment analysis the prediction values improved 14.43% of accuracy.

From review of existing research work, it is found that there is a considerable amount of research gap is available in Information extraction from Tamil gynaecological dataset using machine learning algorithms. The implementation of machine learning methods in Tamil Gynaecological domain data to extract the useful information using ontology is distinct.

3.1. Objectives

The main objective of this research work is to extract useful information from unstructured data using Machine learning algorithm with Ontology representation.

- To classify the dataset as predefined categories using machine learning algorithm
- To obtain the needed information according to the user query dataset using Ontological structure.
- To design the Information extraction framework to project the extracted information in structured format.

4. IE SYSTEM DESIGN

The Information extraction system overall design is given in the Figure 2. The steps involved in implementation of IE system are keyword extraction from query, classify the query based on classified datasets, relation extraction using ontology, fill the extracted data in IE framework.

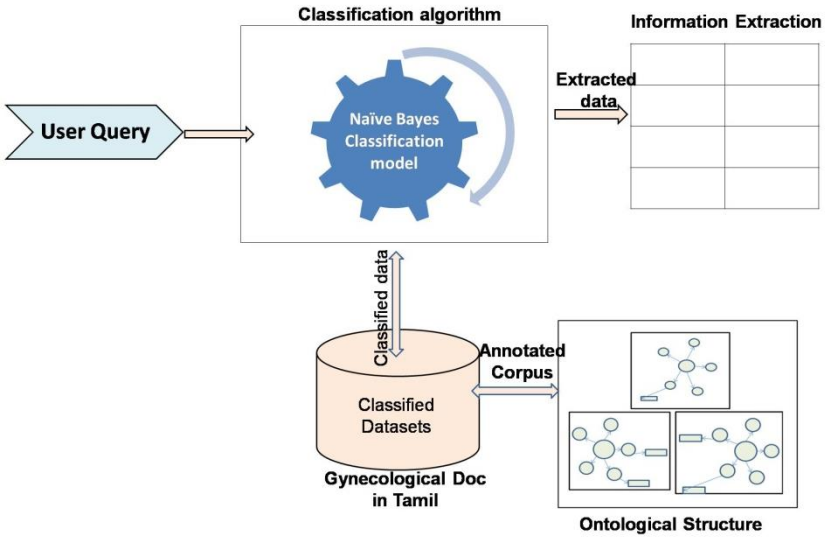

Figure 2. IE system design

4.1. Keyword extraction from query

From the user query the keyword are identified using list of keywords related to the particular issues. Except from keywords other all the details are trimmed in the user query. For example, the query is கருப்பைவாய் புற்றுநோய் ஏற்படுவதற்கான காரணங்கள்?. (Reasons for PCOS) The keywords from the query கருப்பைவாய் புற்றுநோய் & காரணங்கள் (PCOS & Reason) are extracted. These keywords are given as input to get the related data from the classified datasets in corpus.

4.3. Classify and retrieve the elements using machine learning model

The machine learning classification model naïve bayes uses the extracted keywords for classify and retrieve the needed information from corpus. The given query is, கருப்பைவாய் புற்றுநோய் ஏற்படுவதற்கான காரணங்கள்?. (Karuppaivaay puRRunhOykkaaNa kaaraNangaL).
To classify the given query into the appropriate entity list, it will calculate the total number of queries related in each entity. Then it will calculate,

$$P\ (karuppai\ |\text{Karuppai puRRunhOykkaaNa kaaraNangaL})$$

$$P\ (pirappuRuppu\ |\text{Karuppai puRRunhOykkaaNa kaaraNangaL})$$

The probability of given query with all the entity is to be calculated.

P ($karuppai$ |Karuppai puRRunhOykkaaNa
kaaraNangaL = $p(karuppai\ |karuppai)\ \times$
$p(karuppai\ |puRRunhOykkaaNa)\ \times$
$p(karuppai\ |kaaraNangaL)$.

The probability of selected entity and each word in query is calculated. The entity and word, which produce highest value of probability, is selected as most optimal entity of current query. Here,

$p(karuppai\ |karuppai) = 0.98$
$p(karuppai\ |puRRunhOykkaaNa) = 0.24$
$p(karuppai\ |kaaraNangaL) = 0.46$

So, it found that $p(karuppai\ |karuppai)$ is the optimally selected entity for the required query.

4.2. Extraction of needed relations and entities from Ontology

The extracted keywords are used as keys to search in the ontological structure, where the keyword is available with attributes and relation with other entities. The place and the relations with other entities of given keyword is shown in figure 3.

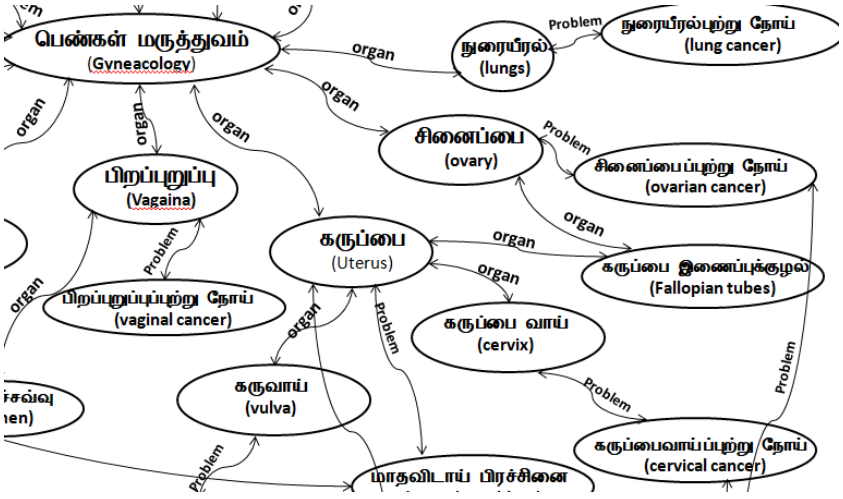

Figure 3. Ontology structure for Karuppaivaay puRRunHoY

The ontology structure is stored as OWL (Ontology Web Language) or RDF (Resource description framework). From the list of keywords in ontology the target elements are identified using machine learning method. Each entities has its relation ids, which it have relations to other entities. The following entity relation table will give details about it. Each entity has its unique entity_id, root_id, child_id1, child_id2, child_id3. By using these entity relations the ontology will be generated for the current selected

entity. In this research work the ontology is implemented using linked graphs with elements. The links will represent to the child entity using attributes of the current entity. For example, the entity Karuppai ➔ is derived from pengaL entity via ➔ ulluRuppukaL attribute.

E.No	Entity	R_ID	C_ID 1	C_ID 2	C_ID 3
456	karuppai		468		
457	Pengal	468			
458	karuppai vEkkam		457		
459	karuppai adiyiRakkam		457		
460	karuppai puRRunhOy		457	469	
461	karuppai agappadala nOy		457		
462	karuppai suvar pirassiNaikaL		457		
463	karuppai kattikaL		457		
464	karuppai nhErkkattikaL		457		
465	P.C.O.D		457		
466	P.C.O.S		457		
467	karuppai vAyppuN		457		
468	ULLuRuppu		457		
469	puRRunOy		423		

Figure.4 Entity relation table

4.3. User Interface of current model

The user interface for the IE framework is designed in scripting languages. The user can give their queries in local language Tamil as follows. Gathered keywords from the query is processed and classified the datasets to retrieve the needed information for the given query. User interface for query processing is shown in Figure 4 & 5.

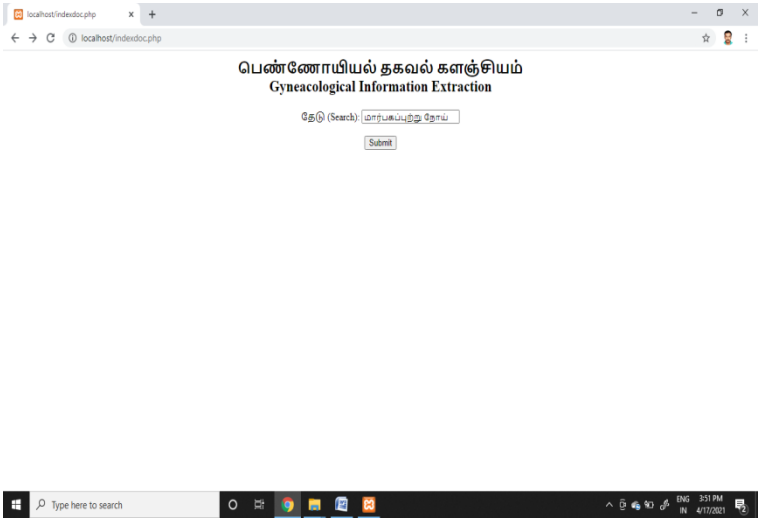

Figure.7 Query Interface for IE

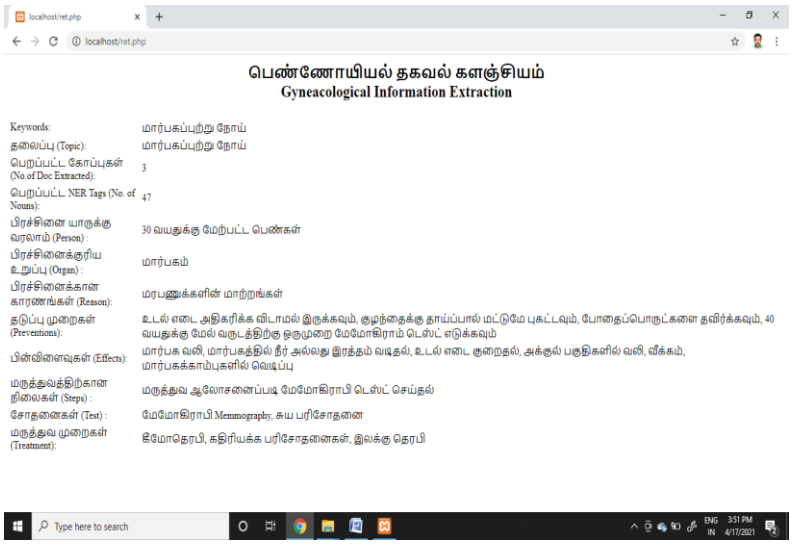

Figure.8 Extraction of data using IE

The sample query has been given and it is verified that the relevant information is retrieved as correctly.

4.4. Evaluation and discussion

To check the correctness of the IE framework results the Precision and Recall method is used. The precision and recall evaluation method is based on the following confusion matrix.

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

Table.1 Confusion Matrix

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

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

And, finally the F-Score is calculated as follows,

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

4.4.1. Traing phase

The classification and retrieval task is splited into two phases training and testing phase. In training phase there are 9 categories of datasets are used to train the naïve bayes model. There are 1635 documents are involved in training phase. The accuracy of training phase is given in Table 2.

Topic	No. of Doc	Accuracy
		Naïve Bayes
ஆரோக்கியபாடம் (Women Health Education)	211	86.2559242
வெள்ளைப்படுதல் (White discharge)	167	84.4311377
சினைப்பை புற்றுநோய் (Ovarian Cancer)	127	72.4409449
கருப்பை நீர்க்கட்டிகள் (PCOS/PCOD)	231	84.8484848
பிறப்புறுப்பு புற்றுநோய் (Vaginal Cancer)	195	78.974359
கருப்பைவாய் புற்றுநோய் (Cervical Cancer)	168	82.7380952
மார்பகப்புற்றுநோய் (Breast Cancer)	206	81.5533981
தைராய்டு (Thyroid)	153	79.7385621
மாதவிடாய் பிரச்சினை (Mentrual Problems)	177	89.2655367
Overall Accuracy	**1635**	**82.6911315**

Table 2. Accuracy of training phase of Naïve Bayes model

The confusion matrix for the naïve bayes is shown in Table 3. In the confusion matrix, it is found that the number of documents classified as True positive.

Topics	No. of Doc		Confusion Matrix	Predicted									
				0	1	2	3	4	5	6	7	8	Accuracy
ஆரோக்கியபாடம் (Women Health Education)	211	Actual	0	195	13	3	0	0	0	0	0	0	0.863
வெள்ளைப்படுதல் (White discharge)	167		1	1	158	1	0	0	0	0	6	1	0.844
சினைப்பை புற்றுநோய் (Ovarian Cancer)	127		2	2	4	117	0	0	0	1	1	2	0.724
கருப்பை நீர்க்கட்டிகள் (PCOS/PCOD)	231		3	3	0	1	217	5	2	1	1	1	0.848
பிறப்புறுப்பு புற்றுநோய் (Vaginal Cancer)	195		4	0	0	3	0	185	2	2	1	2	0.790
கருப்பைவாய் புற்றுநோய் (Cervical Cancer)	168		5	0	0	2	1	4	160	0	0	1	0.827
மார்பகப்புற்றுநோய் (Breast Cancer)	206		6	0	0	0	0	0	0	202	2	2	0.816
தைராய்டு (Thyroid)	153		7	0	0	0	1	0	0	0	149	3	0.797
மாதவிடாய் பிரச்சினை (Mentrual Problems)	177		8	0	0	0	0	0	0	0	2	175	0.893
Total	**1635**								**Overall Accuracy**				**0.827**

Table 3. Confusion Matrix

Based on the confusion matrix and classification task results the naïve bayes model given good results (82%) in training phase.

4.4.2 Testing phase

To test the model using real time datasets, sample questions are collected from real life users (friends, classmates, colleagues). A number of 57 unique keywords are collected as test datasets from the real life users. The list of keywords from questions is given in the table 4.

By implementing the Naïve bayes algorithm with the Query datasets, there are two unknown keywords are collected from user queries, கர்ப்பம் தரித்தலில் பிரச்சினை (Problems in Pregnancy) and சிறுநீரகத்தொற்று (Urinary Infection). To test the accurate performance of the selected model with these unknown keywords the datasets are used to test. The accuracy of the model for test datasets is given below.

Issues (Keywords)	No. of Queries	Accuracy
ஆரோக்கியபாடம் (Women Health Education)	1	0.8484
வெள்ளைப்படுதல் (White discharge)	5	0.7908
சினைப்பை புற்றுநோய் (Ovarian Cancer)	8	0.8263
கருப்பை நீர்க்கட்டிகள் (PCOS/PCOD)	7	0.7589
பிறப்புறுப்பு புற்றுநோய் (Vaginal Cancer)	1	0.7897
கருப்பைவாய் புற்றுநோய் (Cervical Cancer)	4	0.8431
மார்பகப்புற்றுநோய் (Breast Cancer)	10	0.8625
தைராய்டு (Thyroid)	6	0.8926
மாதவிடாய் பிரச்சினை (Mentrual Problems)	8	0.8155
கர்ப்பம் தரித்தலில் பிரச்சினை (Problems in Pregnancy)	4	0.1257
சிறுநீரகத்தொற்று (Urinary Infection)	3	0.5896
Testing Accuracy	57	**0.74028**

Table 4. Accuracy of Testing phase

The accuracy of testing dataset is reduced as 74%. Because there are two unknown dataset queries added in testing datasets when compared with training dataset. The Precision, recall and F1-Score values of training and testing datasets for naïve bayes model are given in Table 5.

Datasets	Precision	Recall	F1 Score
Training	0.9713	0.8103	0.8835
Testing	0.8417	0.6854	0.7593

Table 5. F1 Score of datasets

The naïve bayes classification model performed well for the selected gynaecological domain in Tamil language. It gives good accuracy on training (82%) and testing (74%).

5. CONCLUSION

The Information extraction framework created successfully and it works effectively for the selected domain text using machine learning methods. For the training phase there are 1635 documents related to women health issues were stored as entity corpus to train the model. From real life end user it collected about 57 queries for testing. Based on ontology, entities and attributes the final information extraction framework is created. From user query text the entities and relations are classified by using naïve bayes classifier. Each entity in corpus list has its node details of ontological location and representation. Based on ontology the framework details to be filled automatically. To evaluate the correctness of the model the F-score, precision and recall method is used. It is found that the IE framework yields good results (F-score = 0.88 & 0.75) with training and testing datasets. In future the IE system will be implemented for Gynaecological IE system with audio visual properties of the Gynaecological domain data in Tamil language.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] R. Sankaravelauthan, A. Kumar M, Visual Onto-thesaurus for Tamil, Language in India, 17 (2017), 5.

[2] P. Mathur, K. Sathishkumar, M. Chaturvedi, et al. Cancer Statistics, 2020: Report From National Cancer Registry Programme, India, JCO Glob. Oncol. 6 (2020), 1063–1075.

[3] V. Jose, V.P. Jagathy Raj, S.K. George, Ontology-Based Information Extraction Framework for Academic Knowledge Repository, in: X.-S. Yang, S. Sherratt, N. Dey, A. Joshi (Eds.), Proceedings of Fifth International Congress on Information and Communication Technology, Springer Singapore, Singapore, 2021: pp. 73–80.

[4] C. Chantrapornchai, A. Tunsakul, Information Extraction on Tourism Domain using SpaCy and BERT, ECTI Trans. Computer Inform. Technol. 15 (2021), 108-122.

[5] S. Bharathi, A. Geetha, R. Sathyanarayanan, Sentiment Analysis of Twitter and RSS News Feeds and Its Impact on Stock Market Prediction, Int. J. Intell. Eng. Syst. 10(6) (2017), 68-77.

[6] P. Zhou, N. El-Gohary, Ontology- based automated information extraction from building energy conservation codes, Autom. Construct. 74 (2017), 103-117.

[7] S. Jusoh, A. Awajan, N. Obeid, The Use of Ontology in Clinical Information Extraction, J. Phys.: Conf. Ser. 1529 (2020), 052083.

[8] D. Thenmozhi, C. Aravindan, Ontology-based Tamil–English cross-lingual information retrieval system, Sādhanā. 43 (2018), 157.

[9] S. Bharathi, A. Geetha, Sentiment Analysis of Effective Stock Market Prediction, Int. J. Intell. Eng. Syst. 10 (2017), 146–154.

[10] M. Rajasekar, A. Geetha, Comparison of Machine learning methods for Tamil Morphological Analyzer, Intelligent Sustainable Systems, Proceedings of ICISS 2021, (2021), ISBN 978-981-16-2421-6.