# An Improved Approach of Unstructured Text Document Classification Using Predetermined Text Model and Probability Technique

Sreedhar Kumar S, Syed Thouheed Ahmed,Mercy Flora P,Hemanth L S, Aishwarya J, Rahul GopalNaik,Ayesha Fathima

Department of Computer Science Engineering,
Dr.T. Thimmaiah Institute of Technology,
Kolar Gold Fields – 563122, Karnataka India

syed@drttit.edu.in

**Abstract.**Document classification is the task to split the document set into distinct highly relative classes or groups based on nature of the document contents.Here**,** an improved approach of document classification called keyword-based document classification (KBDC) is introduced. It focuses on splitting the unstructured text document set into K number of dissimilar classes based on K predetermined keywords text models by improved probability technique. This new system comprises of the following stages. Namely, pre-processing, classification and classifier stage respectively. Initial, the proposed system (KBDC) recognizes all the immaterial existing contents in the input text document through constructed Predetermined Irrelevant Text Pattern Model (PITPM). Next, it divides the pre-processed document set into 'K' different groups or classes by K number of Pre-determined Keyword Text Pattern Models (PKTPM) through probability technique, where K denotes the number of groups or classes or models. Finally, the KBDC system classifies the trial test text document without any class label that belongs to either of the existing group based on the K different class models (PKTPs). Experimentation results show that the KBDC is appropriate to split and identifies the unstructured text document set into K distinct extremely comparative classes.

***Keywords*:**Classification, Classifier, Keyword Based Document Classification (KBDC), Predetermined Irrelevant Text Model, Probability Technique, Predetermined Keyword Text Pattern Model (PKTPM)

## 1 Introduction

Document classification is a procedure that involves the classifying of text documents to their respective set of classes which are predefined with labels [6, 7]. It is a supervised learning approach in which training set of documents $\{D_1, D_2 \ldots D_n\}$ labeled with classes from {1…...m} are used to construct a model and predict the class label

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2304041

of a new incoming document based on the training model. The document classification in machine learning is divided into two types i.e. unsupervised learning and supervised learning. In unsupervised documents without labels are taken to identify for hidden structures by methods of clustering.

Whereas, in supervised learning the document classification is done using various algorithms that include linear regression, support vector machine (SVM), naïve Bayes etc. A single labeled document is one which has a single class and when assigned with more than one class it is called multilabel classification. Binary classification is type of classification which helps to determine whether chosen document fits into a particular class or not and it serves as an example for single label classification. Generally, text classification is based on some of the stages like preprocessing, feature extraction, dimensionality reduction, classification, validation and verification etc. The first stage is the pre process stage where the text documents are split to form tokens. These tokens may be words or phrases. All these split tokens are preprocessed to remove all the unwanted stop words i.e. the removal of all the unwanted words from the input document so, as to reduce the size of the document to later classify the same document easily. It also removes the common insignificant words that occur in the document like 'but', 'a', 'the', 'an' etc. Stemming is one of the techniques applied where it drops all the unwanted character mostly suffix to reduce the word to its root form.

All The collected text documents are associated with a specific tag that are inputs to the classifier and this process is known as indexing. Here, the documents are represented with features or terms. Generally, these terms may be various phrases or words. The occurrence of the word in the document is based on weight that is assigned to that word in the document. Assigning weights to the words is termed binary indexing where the weight is 1 if the word is present in the document or else 0. The common methodology used is TF-IDF. The TF-IDF stands for term frequency - inverse document frequency, it is a method most commonly used in text mining for filtering the stop words in various fields that includes text classification and summarization. The disadvantage in this methodology used is the score of semantics is not taken into consideration[7]. The documents to be classified can be of various formats like a text, image, music etc. All these documents possess its special classification problems. There are also various probabilistic techniques for document classification.

## 2 Literature Survey

A literature survey is brief summary of previous research work on a particular topic [6] [8][9]. The work should enumerate, summarize and objectively evaluate the previous research. The author JavedMostafa et.al [1] reported that the document classifiers plays an intermediate role in filtering Multi labeled systems using supervised learning methods which results in evaluating accuracy when the work is compared with the baseline system that use the classification result given by the humans. Neural Network along with the Filtering System is one of the concepts in supervised learning method used in this research. The work presented in this paper consists of two phases.

Finally, the results show that high degree of accuracy was achieved by this neural network with filtering. PoojaBolaj et.al [2] presented the efficient classification of the document that is the language Marathi based on the various supervised technique and ontology-based classification. The techniques used shows high accuracy and better time efficiency. It has 3 phases that is pre-processing, feature extraction, supervised learning and ontology-based classification. In this system the given input is Marathi language document which undergoes pre-processing stage including input validation, tokenization, stop word removal, stemming and morphological analysis. Then the classification of Marathi language is done.

Syed MuzamilBasha et.al [3] presented that larger number of machine learning techniques is used in classification of document. However, most of the techniques are based on supervised learning which requires either pre-training or human intervention for classification. The main aim is to discover the best Supervised Machine Learning Algorithms in Document classification using Document Term Vector representation method and to perform huge data organization that saves lot of user time and helps in analyzing customer feedback and compare the result to evaluate better accuracy. Here, Term frequency is used as a feature in Document Classification and Analytical platform KNIME is used with supervised learning algorithms. The result states that SVM is the best-supervised machine learning algorithm used in the area of document classification, compared to DT and KNN. Kabita.Upendra Singh et.al [4] has reported that the with the fast growth in the web has rendered the document classification by humans a challenging task which has given opportunity to various other techniques like data mining, natural language processing (NLP), and machine learning for automatic classification of textual documents.

Hence, if the documents which are taken as input are in the digitalized form then the process of automated text classification is very efficient and useful if the documents are in large quantity. Sreedhar Kumar S et.al [5] presented a system where online commands is taken from various media as input and preprocessed to nullify the irrelevant words present in that command based on the constructed trained model of irrelevant words , next the preprocessed document is tested with the trained model to check whether the command is truly positive or negative based on the sentimental analysis technique. [10][11] The framework comprises of two stages, first stage is recognition of sentiment and second is relative scoring for every element. The purpose of this study is to analyses the performance of sentiment analysis in the terms of accuracy, precision and recall. Finally, it identifies the best algorithm used in this work.[12][13]

## 3 KBDC System

In the following segment, detailed proceedings concerning a newly developed system called (KBDC) is discussed. This system mainly aims to preprocess the unstructured documents from various sources of internet and classify those raw documents to their respective classes by using probability techniques and supervised concepts. The pro-

posed KBDC system comprises of three stages, first stage is the pre-processing of the text document, next classification(training) followed by the classifier stage. Firstly, the suggested system implies pre-processing of the unstructured input text documents set that is often collected from various sources of internet later to count the number of words presents in that document set. Next based on the Predetermined Irrelevant Text Pattern Model (PITPM) the documents are reduced by removing all the irrelevant words to obtain a preprocessed document. In the second stage, it splits the reconstructed text documents set into K distinct classes based on K distinct models namely PKTPM-e (Engineering), PKTMP-m (Medical), PKTMP-l (Law), PKTPM-b (Business) through improved probability technique. In the final stage, the KBDC system recognizes and classifies the recent sampled unstructured text document to either of the four class i.e. (medical, engineering, business or law) based on trained document set. The distinct phases involved in this propound KBDC system are depicted in the Fig 1.

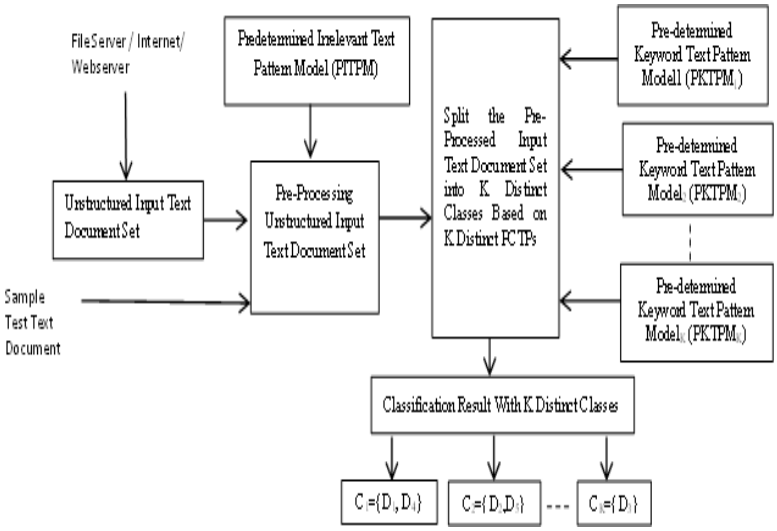


**Fig.1.**Architectural depiction of Proposed KBDC Approach.

### 3.1 Pre-Processing Stage

Pre-processing stage is the first stage in the propound system Here, KBDC system primarily concentrates on reducing the number of irrelevant word(keyword) from the raw documents taken from various sources of internet. The raw documents are generally unstructured with no labels. These documents are pre-processed through (PITPM) model by identifying the irrelevant word like article, prepositions, conjunction etc. In the beginning, it identifies the immaterial existing words in the input document set $D = D_i$ for $i = 1,2,3,..,n$ , where $D_i$ represents the $i^{th}$ document in the set of documents $D$ and $n$ describes the number of documents in the $D$ . In the next stage, the collected

document is pre-processed with the already constructed model (PITPM), to remove all the immaterial words using simple technique of word matching to obtain a pre-processed document as $\overline{D} = \overline{D}_i$ for $i = 1,2,3,..,n$ where $\overline{D}_i$ denotes the $i^{th}$ document in the reconstructed document set $\overline{D}_i$ with n distinct documents.

### 3.2 Classification Stage

This stage, the KBDC system divides reconstructed text document set $\overline{D}$ into 'K' dissimilar classes namely medical (m), engineering (e), business (b) and law (l) viz. 'K' Predetermined Keyword Text Pattern Models (PKTPM) and probability technique. The PKTPM is containing four different predetermined classes i.e. Medical, Engineering, Business and Law containing limited number of particular class key words. Two steps are followed in this stage. Firstly, it calculates the probability of class keywords likely, Medical keywords P( $\overline{D}_i$ | "Medical keywords")over each individual text document present in the pre-processed document set $\overline{D} = \overline{D}_i$ for $i = 1,2,3,..,n$ where $\overline{D}_i$ denotes the $i^{th}$ text document belongs into pre-processed document set. Where, |m| denotes the number of predetermined keywords in medical class and it is given by the Equation (1) as:

$$P(\overline{D}_i \mid \text{"}medical\ keywords|m|\text{"}) = \left\{\left(\frac{\overline{D}_i \text{ in "m"}}{|m|}\right), \forall\, \overline{D}_i \in \overline{D}\right\} \quad (1)$$

Where,"$m$"denotes the medical class keywords in$i^{th}$text document that belongs to preprocessed document set. finally, it identifies whether $i^{th}$ text document $\overline{D}_i$ in the preprocessed document set $\overline{D}$ belongs to Medical, Engineering, Business or Law class based on calculated results of probability that is given in equation (1). Suppose, the probability of medical class keywords is greater than other class keywords correspondingly the $i^{th}$ document of $\overline{D}_i$ is placed (Marked) into medical class (m) otherwise, the process of testing the document belonging to the set $\overline{D}_i$ is followed with all the other three pre-determined classes to finally classify that documents to the class which is the having the highest probability.

### 3.3 Classifier Stage

In the last stage, following KBDC approach helps to classify the new sample test text document to its respective class based on probability technique. In the initial step, the KBDC identifies the immaterial words in the new sample text document $(d)$ with no class labels based on (PITPM) that is previously discussed in the subsection (3.1) and the result is defined as $(\overline{d})$, where $\overline{d}$ is the reconstructed sample text document. Next, it identifies the reconstructed document $(\overline{d})$ belongs to which of the existing class namely 'm', 'e', 'l' and 'b' by the process of measuring the probability of doc-

ument $\bar{d}$ in the "K" distinct class models namely 'm', 'e', 'l' and 'b' respectively and it defined in the equations (2), (3), (4) and (5) as:

$$P(\bar{d} \mid m) = \frac{(\bar{d} \cap m)}{|m|} \quad (2)$$

$$P(\bar{d} \mid e) = \frac{(\bar{d} \cap e)}{|e|} \quad (3)$$

$$P(\bar{d} \mid l) = \frac{(\bar{d} \cap l)}{|l|} \quad (4)$$

$$P(\bar{d} \mid b) = \frac{(\bar{d} \cap b)}{|b|} \quad (5)$$

Where, $m$ is the predetermined medical class model in the PKTPM, $|m|$ denotes the number of words in the $m$ model, $e$ represents the scheduled engineering class model, $|e|$ describes the size of class model $e$, $l$ indicates the fixed law class model, $|l|$ represents the number of arguments in the $l$ class model, $b$ represents the keyword present in business class model and $|b|$ signifies the quantity of words in the $b$ class model. Lastly, the KBDC identifies the document $\bar{d}$ to its highly relative class based on the higher probability of prearranged class model set $Z$ with $K$ class models for $Z_i$ $i = 1,2,3,..,n$ and it is defined in the equation (6) as:

$$Max\left\{ P(\bar{d} \mid Z_i) = \left\{ \frac{\left(\bar{d} \cap Z_i\right)}{|Z_i|} \right\}, \quad \forall Z_i \in Z \right\} \quad (6)$$

### 3.4 Pseudo Code For Keyword Based Document Classification (KBDC)

**Input**: Unstructured Existing Text document Set
D= $\{D_1, D_2, \ldots, D_n\}$, Recent Text document $(d)$, PITPM, PKTPM.
**Output**: Classified Result of sample document $(\bar{d})$.
**Begin**

1. Unstructured document $D=D_i$ for $i$=1,2,3…...n is pre-processed based on Predetermined Irrelevant Text Model (PITM).

2. Pre-processed existing text document set $\overline{D}=\overline{D}_i$, with $n$ documents is split into four distinct classes i.e. medical, engineering, business and law based on PKTPM and simple probability technique Equations (1).

3. Count the total number of keywords in all the classes (|m|), (|e|), (|b|), (|l|) and the documents of the pre-processed document set $\overline{D}$.

4. Calculate prior probability of all classes over the pre-processed existing text document set $\overline{D}=\overline{D}_i$ using Equation (2-5).

5. Calculates conditional probability of medical, engineering, business and law class over pre-processed sample document $\overline{d}$ based on pre-processed text document set $\overline{D}$=$\overline{D}_i$ using Equation (6).

6. Identify the class which is having highest conditional probability for the recent text document $\overline{d}$, later to classify that document to its respective class.

**End**

## 4 Results and Discussions

The following section describes the complete work along with the experimental result of propound KBDC system which is tested over the existing unstructured document set. For the being of fact-finding, increased number of unstructured text documents with dissimilar size has taken from the internet which is being related to various fields and domains. Here, Table 1 shows the 6 documents which are taken as an example. Table 1 contains the finite number of unstructured text documents $D = D_i$ for $i = 1,2,3,..,n$ and also encompassing the size of each individual document in the $D$. In this stage primarily, the KBDC system firstly pre-processes the collected unstructured text document set D containing n documents. Here, n=6 that are shown in the Table 1 which are pre-processed based on the constructed predetermined irrelevant text pattern model (PITPM) given in the Table 2. This table comprises of a smaller number of stop words likely articles, prepositions and Conjunctions etc. Later to obtain a pre-processed text document set $\overline{D}$ with n documents which is shown in Table 3. Next, the KBDC system partitions the previously pre-processed text document set $\overline{D}$ into four classes i.e. medical, engineering, business and law based on probability of each class keywords that is present in each individual document of the predetermined keyword text pattern model (PKTPM).

**Table 1.**Existing Unstructured Input Text Document**.**

| **Document id** | **Unstructured text document ($D_i$)** | **Number of words in $\|D_i\|$** |
|---|---|---|
| 1 | Medical science is a field of science where doctors and other health professionals uses various methods of dragonizing, and curing a particular illness or a disease. Various surgical methods are used to perform operations and cure all the major disorders. They mostly prefer to use instruments and surgical methods to cure the abnormality rather than using the manual techniques. They often perform surgery to eliminate the affliction from the biopsy. More often, doctors prefer to remove the defect causing tissues from the body. Health professionals uses a wide range of instruments in order to diag- | 125 |

| | | |
|---|---|---|
| | nose a condition, and treat it for the well. Medical instruments ranges from a small test tube to other sophisticated machines. | |
| 2 | Science is a vast area which comprised of many fields of which engineering is a major application. generally, engineering is a field concerning of solving the given problem in technical manner. Generally, scientist and innovators are the peoples who innovate things that bring up advancement in the human life. whereas, engineers are those who bring out those innovations to use for the betterment of the humans. In his book, "Disturbing the Universe" , physicist Freeman Dyson wrote, "A good scientist is a person with original ideas. A good engineer is a person who makes a design that works with as few original ideas as possible. | 121 |
| 3 | In today's world, business plays a vital role in making money and one's own living. Simply put, it is "any activity or enterprise entered into for profit. Generally, business can range from a small street peddler to a multinational company. It does not need to have a partnership, corporation or a company etc. owner of the business will be responsible for the entire happening of the agency he will responsible for depts occurred in the business. If the business acquires debts, tax professional and other creditors are liable to go after the possessions of the owner. The term is also often used colloquially (but not by lawyers or by public officials) to refer to a company. A business structure is very difficult to be established and developed but if developed it provides more security to the owner. The owner is taxed for all the profits that he has incurred from the business. | 150 |
| 4 | Business plan is a structured outline the business happenings a plan which includes various blue prints like how the plan must be executed, in what time interval and also predicts the rough output. Generally, it describes the overall nature of the organization, the strategies involved to reach the target, the financial status of the organization etc. these plans majorly servers a blue print for the entirety of the organization goals. Document format of these plans are required to claim loans in banks. Business plans are decision-making tools. | 99 |
| 5 | Business idea, where idea acts a pyramid for the entire business, a business is mainly concerned on a particular commodity or a product that brings out financial gains for the owner. A promising business includes the characteristics like innovative, unique, problem solving and profitable. For a business innovation means creating new ideas through various research and development or improving the existing services. Business idea should concentrate on solving the existing problems in the market in order to become a global leader. Businesses that are focused on innovations are likely to be very efficient, cost effective and productive. | 107 |
| | System of rules mainly created through social or governmental institutions is commonly defines as law. It has been variously described as a science and the art of justice. Law can be divided into two types | |

| | | |
|---|---|---|
| 6 | namely, public and the private law. public law concerns government, society. Private law concerns the disputes between two individual and organizations for property, contacts, delicts. Legal systems vary between countries based on their jurisdictions. Law raises important and complex issues concerning equality for all fairness and justice. | 93 |

Table 1 manifests the constructed (PKTPM) that includes the finite number of class keywords of the four classes respectively. Finally, the KBDC system identifies the first document belongs into medical (m) class and similarly all other documents are identified to their respective classes based on the mathematical results obtained from the probability techniques and various supervised concepts. Table 5, manifests the following outcome. This result shows that the suggested KBDC system has correctly trained the document set$\overline{D}$ to later identify their respective classes.
Finally, the KBDC system is examined with sample recent text documents $(d)$ witch no class labels to check as it fits into which of these class i.e. (m, e, b, l) based on existing probability technique and already trained text document set which is identified in the classification stage. Table 3 shows how to pre-process recent sample document to obtain prepressed recent text document $(\overline{d})$. Following this procedure, it computes the prior probability over the trained document set as P($\overline{d}$|m). Later, it calculated the probability of all the four classes called m, e, l and b respectively with the recent preprocessed sample document $(\overline{d})$ are displayed in the Table 6. Finally, the recent document is categorized to its particular class based on conditional probability of all the trained documents and the end categorization results are manifested in the table. Research results shown in the Table 6, acts as a proof that the implemented KBDC system is suitable to train an unstructured document set with no class labels taken from various resources and classify them to their respective classes.

**Table 3**: Pre-processed Input Documents

| **Document id** $\overline{D}_i$ | **Pre-processed document set** $(\overline{D}_i)$ | **Number of words in** $\lvert\overline{D}_i\rvert$ |
|---|---|---|
| 1 | Medical sciencefield doctor health professionals dragonizing curing illness disease surgical operations cure disorders instrument surgical abnormality manual techniques surgery eliminate affliction biopsy defect tissues body condition treat test tube sophisticated machines | 32 |
| 2 | Science vast area fields engineering application generally solving problem technical manner scientist innovators innovate advancement human life innovations betterment disturbing Universe physicist Freeman Dyson original ideas engineer design | 28 |

| | | |
|---|---|---|
| 3 | business money living activity enterprise profit small street peddler multinational company partnership corporation owner agency debts tax professional creditors liable possessions colloquially lawyers public officials structure established developed security profits incurred | 31 |
| 4 | Business plan structured outline happenings blue prints executed time interval rough output nature organization strategies target financial status entirety goals document format claim loans banks decision making tools | 28 |
| 5 | Business idea pyramid commodity product financial gains owner promising business characteristics innovative unique problem solving profitable business innovation creating research development improving existing services concentrate solving existing problems market global leader efficient cost effective productive | 35 |
| 6 | System rules social governmental institutions law science art justice public private public government society Private law disputes individual organizations property contacts delists Legal systems countries jurisdictions complex issues equality fairness justice | 31 |

**Table 2.**Pre-determined Keyword Text Pattern Models (PKTPM) $(Z)$.

| **Engineering (e)** |
|---|
| Engineering, kinematic chains, computer science, electronics, mining, design, optical fibers, construction, electrical circuits, mechanical, engine, vacuum technology, compressor, power, software, research, generators, machine, mechanics, chemical engineering, manufacturing, petroleum, refining, microfabrication, fermentation, civil ,structural ,environmental , surveying, electrical ,Broadcast , chemicals, motor, computer systems, electrical ,circuits, generators, motors, electromagnetic, electromechanical devices, electronic , electronic circuits, optical fibers, optoelectronic , computer systems, telecommunications, instrumentation, controls etc. |

| **Medical (m)** |
|---|
| Medicine, pharmaceutical, drugs, Psychiatric, blood, stethoscope, radiography, diagnose, treatment, vascular, respiratory therapist, transplant, diseases, injection, trauma, doctor, pharmacists, health, healing, prevention, speech therapist, operation theater, hospital, surgery, scanning, neurosurgery, surgeon, research specialists, orthopedic, vaccinations, Neurological, Musculoskeletal, Ophthalmic, Cardiovascular, dentist, Colorectal, Pediatric etc. |

| **LAW (l)** |
|---|
| Lawyer, trail preparation, policy, Law, court, criminals, attorney, legal, judges, litigation, solicitor, risk management, witness, prosecutor, corporate, statutory, custody, common law, contract, barrister, deposition, dispute, counsel, sister in law, criminology, family law, civil law, law firm, legal advice, case law, management, real estate, fraud, justice system, legal, constitution, policy development, protection, memoranda, police, jail, surveillance, resolution, rights, |

| judges etc. |
|---|

| **BUSINESS (b)** |
|---|
| Inventory, business, manufacturing equipment, management, strategic , selling price, operations , profit, service, retail fixtures, banks, marketing, brokerage firms, business plan, credit unions, asset , credit cards, public finance, insurance, selling, companies, investment , business idea, private equity firms, business news, equity funds, human resource, real estate, buying, personal finance, investment, trusts, sovereign, opportunity, wealth funds, production , corporate finance, pension , mutual , accounting, index , hedge , income, stock exchanges, Products, loss, money, cost price, manufacturing, marketing, purchasing etc. |

**Table 3.**Classification Result of Input Document Set.

| **Preprocessed Input Document ID** $(\overline{D}_i)$ | **Probability of medical (m) keywords in $(\overline{D}_i)$ P $(\overline{D}_i \mid$ m)** | **Probability of engineering (e) keywords in $(\overline{D}_i)$ P $(\overline{D}_i \mid$ e)** | **Probability of business (b) keywords in $(\overline{D}_i)$ P $(\overline{D}_i \mid$ b)** | **Probability of law (l) keywords in $(\overline{D}_i)$ P $(\overline{D}_i \mid$ l)** | **Classification Result of Input Document (m/e/b/l)** |
|---|---|---|---|---|---|
| 1 | 0.864 | 0.212 | 0.03 | 0 | 'm' |
| 2 | 0.27 | 0.571 | 0.09 | 0 | 'e' |
| 3 | 0.243 | 0.204 | 0.574 | 0.22 | 'b' |
| 4 | 0.189 | 0.16 | 0.518 | 0.266 | 'b' |
| 5 | 0.10 | 0.148 | 0.18 | 0.77 | 'l' |
| 6 | 0.216 | 0.632 | 0.148 | 0.11 | 'e' |

| Document id | Unstructured recent text document $(d)$ | $\lvert(d)\rvert$ | Preprocessed recent text document $(\overline{d})$ | $\lvert\overline{d}\rvert$ | $P(\overline{d} \mid m)$ | $P(\overline{d} \mid e$ | $P(\overline{d} \mid$ | $P(\overline{d}$ | Classification Result (m/e/b/l) |
|---|---|---|---|---|---|---|---|---|---|
| 7 | Medicine is a part of science that involves diagnosing and curing of the diseases. Medical encompass various | 86 | Medicine science diagnosing curing diseases Medical fields bio medical genetics dermatology cell biology | 24 | 0.648 | 0.163 | 0.081 | 0 | 'm' |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | fields like bio medical, genetics, dermatology, cell biology, ENT, orthopedic etc. There are numerous specializations in the field of medicine. This field mainly concerns to diagnose the illness and find the correct methods to eradicate the illness. Contemporary medicine applies to diagnose, treat and prevent various diseases and injuries typically through various method of surgery, therapies etc. | | ENT orthopedic specializations illness eradicate illness Contemporary treat injuries surgery therapies | | | | | | |

## 5 Conclusion

Text document classification has been a part of text mining. Here, an improved approach of document classification called keyword-based document classification

(KBDC) is introduced. This new system mainly concerns on categorizing the unstructured documents taken from various sources of internet into either of the four distinct classes i.e. medical, engineering, business and law. Finally, it classifies the sample document to its respective classes with predefined labels using improved simple probability technique. Primarily, it searches the inappropriate arguments in the formless text document through constructed predetermined irrelevant text pattern model (PITPM) and then eliminates those unwanted words like articles, prepositions, conjunctions etc. Next, it recognizes four well defined classes namely medical (m) engineering (e), business (b) and law (l) by prearranged model (PKTPM) and probability techniques over the preprocessed document set. Lastly, the sample recent text document with no class label is affiliated to its respective class based on keyword matching method and already preprocessed text document set

## References


1. JavidMostafa, Wai Lam, "Automatic classification using supervised learning in a medical document filtering application", journal information processing & management (Elsevier), Vol. 36, No.3, pp. 415-444. 2019.
2. PoojaBolaj, SharvariGovilkar, "Text classification Marathi documents using supervised learning", International Journal of computer applications (0975-8887), Vol.155, No.8, pp.1-3, DOI:10,5120/ijca2016912374, 2016.
3. Syed MuzamilBasha, Ambeshwar Kumar, Robbi Rahim, "Comparative study on performance of document classification using Supervised Machine Learning: KNIME", Australian Journal of emerging technologies and society, Vol.10, pp.148-153, 2019.
4. Upendra Singh, SaqibHasan, "Survey paper on document classification and classifiers", International journal of computer science and technology, Vol.3, No.2, pp.83-87, 2015.
5. Sreedhar Kumar S, Syed Thouheed Ahmed, NishaBhai, Vinutha B A, "Type of Supervised Text Classification System for Unstructured Text Comments using Probability Theory Technique", International Journal of Recent Technology and Engineering (IJRTE), Vol.8, No.10,DOI:10.35940/ijrte. B1158.0982S1019, 2019.
6. Gunashree, M., Ahmed, S. T., Sindhuja, M., Bhumika, P., Anusha, B., &Ishwarya, B. (2020). A New Approach of Multilevel Unsupervised Clustering for Detecting Replication Level in Large Image Set. *Procedia Computer Science*, *171*, 1624-1633.https://doi.org/10.1016/j.procs.2020.04.174
7. S. Sreedhar kumar, and M. Madheswaran, "A Brief Survey of Unsupervised Agglomerative Hierarchical Clustering Schemes," International Journal of Engineering & Technology, vol. 8, no. 1, pp. 29-37, 2019.
8. S. Sreedhar kumar, M. Madheswaran, R. Ravi, "Inherent Approach of Medical Image Pixels Classification Using an Improved Agglomerative Clustering Technique", Research Journal of Biotechnology, vol. 12, no. S2, pp. 115-124, 2017.
9. Kumar, S.S., Ahmed, S.T., Vigneshwaran, P. et al. Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets. J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-02487-w
10. Ahmed, S.T., Sankar, S. & Sandhya, M. Multi-objective optimal medical data informatics standardization and processing technique for telemedicine via machine learning ap-

proach. J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-02016-9

11. J. Dafni Rose, K. Vijayakumar and S. Sakthivel, “Students performance analysis system using cumulative predictor algorithm”, Int. J. Reasoning-based Intelligent Systems, Vol. 11, No. 2, 2019.
12. Vijayakumar. K, Nawaz Sherif. T, Gokulnath.S, “Automated Risk Identification using Glove algorithm in Cloud Based Development Environments”, International Journal of Pure and Applied Mathematics Volume 117 No. 16 2017.