ISSN: 2278-4632
Vol-13, Issue-10, October 2023

Some Aspects of Statistical and Concept Based Term Weighting Approaches for Text Categorization

K. SandeepuKumar [1], M. Mathan Kumar [2], Y. Jahnavi [3*]

[1] Dept of Computer Science, Gayatri Vidya Parishad College for Degree and P.G Courses(A), Rushikonda, Visakhapatnam, Andhra Pradesh, India.

[2] Department of Computer Science & Engineering, Karpaga Vinayaga College of Engineering and Technology, Chennai, Tamil Nadu, India.

[3] Dept of Computer Science, Dr V S Krishna Govt Degree and PG College (Autonomous), AU-TDR Hub, Visakhapatnam, Andhra Pradesh, India.

Abstract: Data is abundant and is significant in the knowledge-based system and there is a dire need for knowledge access from the availability of data in the digitalized text document. The proliferation of documents in the information age is staggering. The enormous amount of digitized text has created the requirement for Machine Learning and text mining techniques. Text Mining is the approach of investigating and retrieving desirable knowledge from a set of documents using Information Retrieval, Machine Learning, Data Science and Natural Language Processing techniques. It is simpler to comprehend from grouped text archives. Although this problem has developed in extensive research, pattern extraction is still an open field. There is research that has stated developments and orientation. This paper presents a thoughtful investigation on preprocessing techniques, term weighting algorithms, classification and clustering algorithms, pattern discovery, domain ontology-based frameworks for Natural Language Processing, Machine Learning, and summarization techniques. In augmentation, several productive applications of text mining are reviewed.

Index Terms- Natural Language Processing, Semi structured documents, Unstructured documents, Preprocessing, Term Weighting.

I. INTRODUCTION

Text Mining relates to the invention of non-trivial, formerly unidentified and possibly functional information from an enormous accumulation of text data sets [1][3]. The idea of natural language processing and text mining is data mining on unstructured data. As most of the data is available in unstructured format, the techniques of Natural Language Processing and Text Mining play most important role. Most methods in text mining have been employed to extracting structured datasets, called intermediary forms [13-16]. Text Mining is the most important today, most information is available in text form. The various intermediate representations of text are word, phrase, pattern, concept, paragraph, and document. Any conventional data mining techniques can be applied to these intermediary forms. Text categorization is the function of inevitably framing a set of documents into groups such as concepts. Since the strategy has transformed from a conventional set of static text to a dynamic text stream, where the text stream is a consequence of sequentially arranged documents, the extraction model should not only concentrate on terms of statistical information such as frequency, but also on concepts [2].

Automatic topic extraction is the procedure of identifying significant concepts that meet the characteristic of pervasiveness, whereas in news documents it is the procedure of identifying hot concepts that meet both the characteristics of topicality and pervasiveness [4]. Many applications can also be benefited from the extracted knowledge. Still, finding functional and useful patterns is still an open problem. In this article, we introduce various approaches for obtaining effective patterns and text mining applications.

II. PREPROCESSING

In view of the fact that it is upscale to locate all the terms in all the documents, the reduced traceable list has to be determined by stop word elimination, stemming and pruning of terms that emerge uncommonly [1][5]. The purpose of stop list extermination is to set aside and save the resources of the system [1]. All infrequent terms have to be leftover for experimental computations in the pruning process. The origin behind pruning is that the irregular terms doing not be useful in identifying appropriate clusters [5].

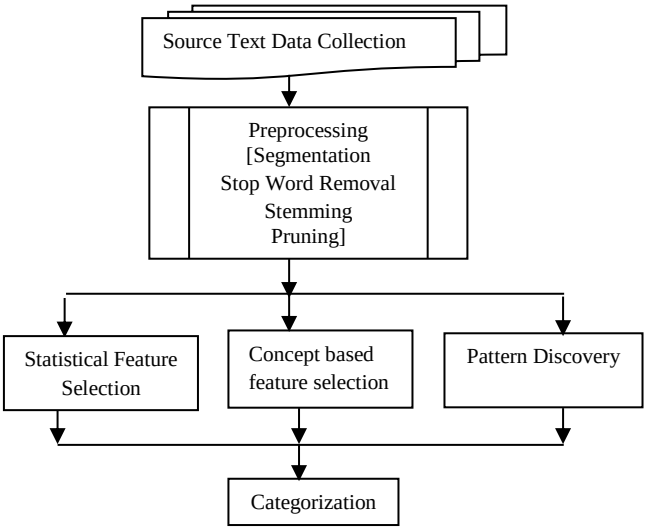

Fig.1 Overall System Architecture.

The general architecture of the system is presented in Fig. 1.

III. TERM WEIGHTING ALGORITHMS

The foremost significant stage in involuntary topic extraction is the process of identifying important terms and sentences. The general idea is to allocate weights to expressions based on their statistical properties. High-weight terms capture ubiquitous information. There are some algorithms that extract terms satisfying the pervasiveness property. There are some algorithms that are simple and easy regardless of the

ISSN: 2278-4632
Vol-13, Issue-10, October 2023

semantic relationship. But the performance of such algorithms is low. Term importance is estimated on the basis of the linguistic, morphological, syntactic and semantic structure of each term rather than term frequency [7][8]. Pattern-based approaches are introduced to conquer the challenges of term-based methodologies. Some of the term weighting algorithms are represented as follows:

(i) Binary Term Weighting:

Term frequency is the computation of the frequently a term appears within a document. It is the value obtained by counting the number of times a term appears within a document.

(ii) Term Frequency and Inverse Document Frequency:

TF-IDF (Term Frequency - Inverse Document Frequency) weights the frequency of a term in a document by a factor that reduces its significance when it occurs in relatively all documents. The term frequency-inverted document frequency for a term t in the document d is calculated by:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i$$

tf is the term frequency and *idf* is the inverse document frequency of term t.

The term frequency could therefore be defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where,

$n_{i,j}$ is the number of occurrences of the considered term in document d_j;

$\sum_k n_{k,j}$ is the sum of the instances of all terms in document d_j.

The inverse document frequency is represented as

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

where,

- $|D|$ is the total number of documents;
- $|\{d : t_i \in d\}|$ is the number of documents that the term t_i occurs.

(iii) Term Frequency and Proportional Document Frequency:

In the concept of TF*PDF, a term is heavily weighted if it appeared in utmost of the channels, and less weighted if it is appeared in a few channels [10]. The weight is calculated as:

$$W_j = \sum_{c=1}^{c=D} |F_{jc}| \exp(n_{jc} / N_c)$$

where,

Normalized term frequency $|F_{jc}| = F_{jc} \Big/ sqrt\left(\sum_{k=1}^{k=K} F_{kc}^2\right)$

W_j = Weight of term *j*;

F_{jc} = Frequency of term *j* in channel *c*;

n_{jc} = Number of documents in channel *c*, where term *j* occurs;

N_c = Total number of documents in channel *c*;

K = Total number of terms in a channel;

D = Number of channels.

*(iv) The Position Weighted TF*PDF:*

The terms in the title are more significant and display the core idea of the news document than the terms in the main text [17].

The position score of term *j* in document *k* in channel *i* is defined as:

$$ps_{ik}(j) = \begin{cases} 3 & j \in \text{the title of the k}^{\text{th}}\text{document in channel i} \\ 1 & j \notin \text{the title of the k}^{\text{th}}\text{document in channel i} \end{cases}$$

The total term position weight will be the sum of the average weight from each channel as follows:

$$pw(j) = \sum_{i=1}^{|C|} \sum_{k=1}^{|N_c|} ps_{ik}(j) / N_{jc}$$

Considering the influence of term position, a novel position-weighted TF*PDF was proposed.

$$weight_j = W_j * pw(j)$$

where,

W_j is the TF*PDF weight.

The member with the highest ranking can be selected by ordering the terms using position-weighted TF*PDF.

(v) Combining Term Frequency with Information Theory Functions or Statistic metrics:

There exist diverse unsupervised term weighting algorithms such as binary tf, $tf.idf$ and its variants and supervised term weighting schemes such as $tf.\chi^2$, $tf.ig$ (information gain), $tf.gr$ (gain ration), $tf.OR$ (Odds Ration) etc. A novel supervised term weighting algorithm $tf.rf$ (relevant frequency) was proposed [18]. On the contrary, by adopting an rf scheme, each term is allocated a more suitable weight in terms of diverse categories.

The relevance frequency is stated as:

$$rf = \log\left(2 + \frac{a}{c}\right)$$

where,

a is the terms which appear in the positive category of a term;
c is the terms which appear in the negative category of a term.
The constant value 2 in the *rf* formula is assigned because the base of this logarithmic operation is 2.
This algorithm considers the frequency of only relevant documents.

ISSN: 2278-4632
Vol-13, Issue-10, October 2023

(vi) Statistical Property (Term Distribution) based Term Weight Algorithm

TF-IDF only considers the frequency term. In TF, the term weight is positively associated with their occurrence. A term with a higher occurrence may actually be heavily dispersed. These expressions are used to express the content of a section as a replacement for the entire document. However, the TF algorithm assigns a more term weight to these conditions. It is exiguous to consider only the frequency of terms and give them weight. The thermal mass distribution algorithm allocates weights in accordance with analogous dissemination scope and spread extension.

(vii) FPST: a new term weighting algorithm:

Various algorithms use the diverse features for extracting salient terms. Some features should be weighted accordingly based on the occurrence in a particular time slice. More weight should be given for the terms appearing in the current time slice and disappears in other time periods. Such property was not reflected in other term weighting algorithms. The position, scattering features along with frequency are also vital. The salient novel algorithm has been proposed for extracting important features [9-12].

(viii) CONCEPT –BASED MINING MODELS:

In text mining, it is significant to select important features that furnish to the creation of salient feature extraction algorithms. Primarily, individual document should be zoned and well-defined boundaries for the sentences should be identified. Concepts in sentences are recognized and labeled based on PropBank entries [19]. Then the sentences in the document are represented in the form of a verb argument structure. A sentence can have more than one verb argument structure. This arrangement enables the formation of a composite expression from the meaning of the individual terms in the sentence. Concept mining has been represented in Fig. 3.

(*ix*) *PATTERN BASED APPROACHES:*

Most feature extraction methods have adopted basic term-based approaches. But it suffers from problems of polysemy and synonymy. A large amount of work has been done in the field of information retrieval based on term methods due to its simplicity and their computational efficiency [20-22]. Although term-based methods suffer from problems of polysemy and synonymy, their advantages include efficient computational power as well as advanced term weighting theories. Sequential patterns have gained interest in the text mining community in the last few years. The reasons for using the sequential pattern approach over term approaches in the text mining community are

i. Compared to a word-based term, a sequential pattern-based feature captures the temporal relationships between words and phrases.
ii. A language model based on sequential patterns has more expressive power than word-based approaches.

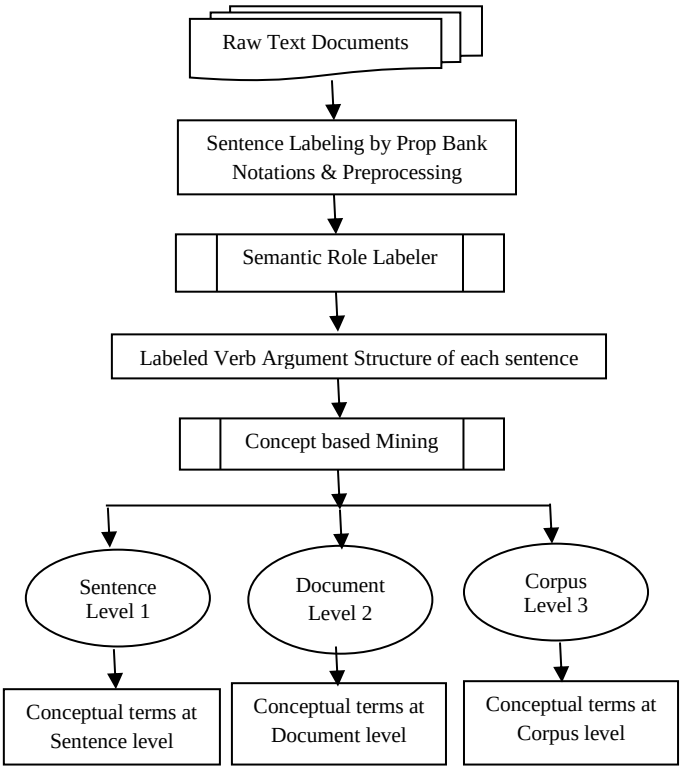

Fig. 3 Conceptual mining

There exist various categorization approaches such as Naive Bayes classifier, Hierarchical classifier, Support Vector Machine, Neural network-based classifier, KNN lazy classifier etc., Clustering techniques such as hierarchical clustering, split clustering, density-based algorithms, self-organizing map algorithm, etc. are also applicable for unstructured data [23-31].

IV. SIMILARITY MEASURES

Similarity measures reflect the degree of proximity among diverse objects, organizing the utmost analogous objects into clusters [20].

EUCLIDEAN DISTANCE MEASURE

It is the utmost regularly used approach to determine the distance among two instances. This measure can be applied for the normalized datasets and k-means algorithm uses this measure as default. It is a conventional metric that computes the dissimilarity among two samples in accordance with the magnitude. $E\left(\vec{t}_a,\vec{t}_b\right)=\sqrt{\left(\sum_{t=1}^{m}\left|w_{t,a}-w_{t,b}\right|^2\right)}$

where,

E the Euclidean distance between two term vectors;
$\vec{t}_a$ Term vector of document d_a;
$\vec{t}_b$ Term vector of document d_b;
m = no of terms;
$w_{t,a}$ and $w_{t,b}$ are term weights.

ISSN: 2278-4632
Vol-13, Issue-10, October 2023

COSINE SIMILARITY

The similarity between the term vectors is calculated by the correlation among the vectors. In this measure, the cosine of the angle between the vectors is calculated.

If the term vectors are normalized to a unit length such as 1 then the two documents are equal.

$$SIM(Doc_i, Doc_j) = \frac{\sum_{t=1}^{m}(Doc_{it} \times Doc_{jt})}{\sqrt{\sum_{t=1}^{m}(Doc_{it})^2} \times \sqrt{\sum_{t=1}^{m}(Doc_{jt})^2}}$$

Doc_{it} is the t^{th} term in the document vector i;

Doc_{jt} is the t^{th} term in the document vector j;

m total number of terms in that document.

JACCARD COEFFICIENT

The main principle in this measure is that the similarity value depends on the common elements. It is also mentioned as the Tanimoto coefficient. Common elements and similarity value are inversely proportional. It is always in the range of -1 to +1.

$$SIM(Doc_i, Doc_j) = \frac{\sum_{t=1}^{m}(Doc_{it} \times Doc_{jt})}{\sum_{t=1}^{m}(Doc_{it}) + \sum_{t=1}^{m}(Doc_{jt}) - \sum_{t=1}^{m}(Doc_{it} \times Doc_{jt})}$$

Doc_{it} is the t^{th} term in the document vector i;

Doc_{jt} is the t^{th} term in the document vector j;

m total number of terms in that document.

This measure is used to find binary differences between two or more objects. It is used in ecological research investigations to determine presence or absence between objects.

EXTENDED JACCARD COEFFICIENT

The binary denotation of jaccard coefficient is enhanced to extended jaccard coefficient with the continuous or discrete objects. It is used for handling the similarity of document data in the text mining.

$$s(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{\|\vec{t}_a\|^2 \|\vec{t}_b\|^2 - \vec{t}_a \vec{t}_b}$$

$\vec{t}_a$ term vector of document d_a;

t_b term vector of document d_b.

It retains the sparsity property of cosine measure which allows discrimination of collinear vectors. It is mostly related to dice coefficient.

PEARSON CORRELATION COEFFICIENT

This method is a measure to determine the similarity of the vectors calculated from the values and their standard deviation. It ranges from [+1, -1]. The Pearson correlation is defined only if both of the standard deviations are finite and both of them are nonzero.

$$SIM(\vec{t}_a, \vec{t}_b) = \frac{m\sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{\left[m\sum_{t=1}^{m} w_{t,a}^2 - TF_a^2\right]\left[m\sum_{t=1}^{m} w_{t,b}^2 - TF_b^2\right]}}$$

$$TF_a = \sum_{t=1}^{m} w_{t,a}$$

$$TF_b = \sum_{t=1}^{m} w_{t,b}$$

$\vec{t}_a$ term vector of document d_a

$\vec{t}_b$ term vector of document d_b

$W_{t,a}$ and $W_{t,b}$ are term weights.

The accuracy in this method increases when data is normalized. This measure yields correct results for any scaling with in an attribute. So, objects with same data with different scales are also calculated.

DICE COEFFICIENT

This measure shortens the denominator of Jaccard coefficient and presents a coefficient 2 in the numerator. As long as the vector values are same, this measure normalization factor remains unchanged. This measure doesn't rely on the common terms.

$$SIM(Doc_i, Doc_j) = \frac{2 \times \sum_{t=1}^{m}(Doc_{it} \times Doc_{jt})}{\sum_{t=1}^{m}(Doc_{it}) + \sum_{t=1}^{m}(Doc_{jt})}$$

Doc_{it} is the t^{th} term in the document vector i

Doc_{jt} is the t^{th} term in the document vector j

m total number of terms in that document.

AVERAGED KULLBACK LIEBLER DIVERGENCE

It measures the distance between two corresponding probability distributions. This measure is also called as "Relative entropy". It is defined as:

$$D_{KL}(\vec{t}_a \| \vec{t}_b) = \sum_{t=1}^{m} w_{t,a} \times \log\left(\frac{w_{t,a}}{w_{t,b}}\right)$$

Since Kullback liebler divergence is not symmetric, averaged kullback liebler divergence is used.

ISSN: 2278-4632
Vol-13, Issue-10, October 2023

$$\pi_1 = \frac{w_{t,a}}{w_{t,a} + w_{t,b}}$$

$$\pi_2 = \frac{w_{t,b}}{w_{t,a} + w_{t,b}}$$

$$w_t = \pi_1 \times w_{t,a} + \pi_2 \times w_{t,b}$$

d_a, d_b = documents
$\vec{t}_a$ term vector of document d_a
$\vec{t}_b$ term vector of document d_b
$T = \{t_1, t_2 \ldots\ldots\ldots t_m\}$ term set
$W_{t,a}$ and $W_{t,b}$ are term weights

The average weighting between two vectors ensure symmetry that is the divergence from document *a* to *b* is same as the divergence from document *b* to *a*.

INTER – PASSAGE SIMILARITY MEASURE

Clustering techniques failed to provide satisfactory results for text documents as the text data is very high dimensional and contains a large number of unique terms in a single document.

As the text document is represented in vector space model the algorithm used to identify the terms is "Bag of Words". This representation is simple and easy but it has some disadvantages.

In this measure the text is divided into segments. The text segments are categorized into three passages. Among those, fixed length passages are to be considered. Both overlapping and non-overlapping passages are used for determine the effect of inter passage similarities on text document clustering.

For a document *d* consisting of *m* terms and assuming window size of *w*, document *d* is segmented into

K windows for non-overlapping text windows.

where, $k = \left(\frac{m}{w}\right) if \quad (m\%w = 0) and \left(\frac{m}{w} + 1\right) if \quad (m\%w > 0)$

K windows for overlapping text windows with size of overlap equal to (w/2)

where,

$$k = \left(\frac{m}{w} - 1\right) if \quad (m\%w = 0) and \left(\frac{m}{w}\right) if \quad (m\%w > 0)$$

m = no. of terms;
d = document;
w = window size;
k = no. of windows.
A window is represented using a feature vector with terms present in passage.

MULTI VIEW POINT-BASED SIMILARITY MEASURE

This similarity measure is an extension of cosine measure. In this measure instead of origin (a single reference point), multiple reference points are considered. Multiple reference points are considered from the outside of the clusters. From this outside point the distances and directions are calculated to the point which is in the cluster. By this measure most similarity assessment is achieved.

Thus, for multiple points the similarity is determined and average of these is to be calculated.

$$MVC = \frac{1}{n - n_r} \sum_{d_h} \cos\left(d_i - d_h, d_j - d_h\right) \left\|d_i - d_h\right\| \left\|d_j - d_h\right\|$$

d_i, d_j, d_h are documents;
n = total no. of documents;
n_r = no. of documents in that cluster.
From d_h the distances and directions to d_i, d_j is calculated.

V. CONCLUSION

This paper introduced preprocessing techniques as well as various categorization methods. From the presented survey, we examine that considerable work has been prepared in the areas of statistical, semantic and natural language processing methods. These can really be functional to an extensive variety of applications. The tendency at the present time indicates to linguistic based and integrated concepts of term-based procedures. However, text mining algorithms and approaches need to be improved.

REFERENCES

[1] Sarker, Iqbal H., "Machine learning: Algorithms, real-world applications and research directions", SN Computer Science, 2.3 (2021): 1-21.

[2] Pérez-Suárez, Airel, José F. Martínez-Trinidad, and Jesús A. Carrasco-Ochoa., "A review of conceptual clustering algorithms", Artificial Intelligence Review, 52.2 (2019): 1267-1296.

[3] Odacioglu, Eyyub Can, and Lihong Zhang., "Text Mining for Rendering Theory: Integrating Topic Modeling to Grounded Theory", Available at SSRN 4141372 (2022).

[4] Ghalandari, Demian Gholipour, and Georgiana Ifrim., "Examining the state-of-the-art in news timeline summarization", arXiv preprint arXiv: 2005.10107 (2020).

[5] Qiang, Jipeng, et al., "Short text topic modeling techniques, applications, and performance: a survey", IEEE Transactions on Knowledge and Data Engineering 34.3 (2020): 1427-1445.

[6] Usai, Antonio, et al., "Knowledge discovery out of text data: a systematic review via text mining", Journal of Knowledge Management (2018).

[7] Usai, Antonio, et al., "Knowledge discovery out of text data: a systematic review via text mining", Journal of Knowledge Management (2018).

[8] Patil, Mrs Neha Sangram., "A survey of pattern discovery methods for Text Mining", IJNRD-International Journal of Novel Research and Development (IJNRD), 3.2 (2018): 13-15.

[9] Jahnavi Yeturu, "FPST: a new term weighting algorithm for long running and short-lived events", Int. J. Data Analysis Techniques and Strategies (Inderscience Publishers), Vol. 7, No. 4, 2015.

[10] Jahnavi Yeturu, "Statistical data mining technique for salient feature extraction", Int. J. Intelligent Systems Technologies and Applications (Inderscience Publishers), Vol. 18, No. 4, 2019.

[11] Jahnavi, Y. and Radhika, Y., "Hot topic extraction based on frequency, position, scatter-ing and topical weight for time sliced news documents", 15th International Conference on Advanced Computing Technologies, ICACT 2013.

[12] Jahnavi Yeturu, "A Cogitate Study on Text Mining", International Journal of Engineering and Advanced Technology, Vol. 1, No. 6, pp. 189-196, 2012.

[13] Jahnavi Yeturu, "Analysis of weather data using various regression algorithms", Int. J. Data Science (Inderscience Publishers), Vol. 4, No. 2, 2019.

[14] Jahnavi, Y., Elango, P., Raja, S.P. et al., "A new algorithm for time series prediction using machine learning models", Evol. Intel. (2022). https://doi.org/10.1007/s12065-022-00710-5).

[15] Yeturu, Jahnavi, et al. "A Novel Ensemble Stacking Classification of Genetic Variations Using Machine Learning Algorithms", International Journal of Image and Graphics (2021): 2350015.

[16] Bhargav, Kanta, S. K. Asiff, and Y. Jahnavi., "An Extensive Study for the Development of Web Pages", Indian Journal of Public Health Research & Development 10.5 (2019).

[17] Yan Gao Jin Liu and PeiXun Ma, "The Hot KeyPhrase Extraction based on TF*PDF", IEEE, 2011.

[18] Deng, Xuelian, et al., "Feature selection for text classification: A review", Multimedia Tools and Applications 78.3 (2019): 3797-3816.

[19] Bakay, Özge, Begüm Avar, and Olcay Taner Yıldız., "Comparing Sense Categorization Between English PropBank and English WordNet", Proceedings of the 10th Global Wordnet Conference. 2019.

[20]. Kowalski, Gerald J., and Mark T. Maybury., "Information storage and retrieval systems: theory and implementation", Vol. 8. Springer Science & Business Media, 2000.

[21]. Ricardo Baeza-Yates, Modern Information Retrival, Pearson Education, 2007.

[22]. Grossman, David A., and Ophir Frieder. Information retrieval: Algorithms and heuristics. Vol. 15. Springer Science & Business Media, 2004.

[23]. Jahnavi. Y, Siva Priya. A. 2018. "An IOT Appliance for Controlling the Fan Speed and Accessing the Temperature through Cloud Technology Using DHT11 Sensor". International Journal on Future Revolution in Computer Science &Amp; Communication Engineering 4 (4):525-28.

[24]. Nagendra, K. V., Y. Jahnavi, and N. Haritha. "A survey on support vector machines and artificial neural network in rainfall forecasting." Int. J. Future Revolut. Comput. Sci. Commun. Eng 3 (2017): 20-24.

[25]. Sukanya et al., "Country location classification on tweets," Indian J. Public Health Res. Dev. 10(5), 890–898 (2019).

[26]. Vijaya, U., Y. Jahnavi, and G. Subba Rao. "Community-Based Health Service for Lexis Gap in Online Health Seekers."

[27]. Jahnavi, Y., V. R. Balasaraswathi, and P. Nagendra Kumar. "Model Building and Heuristic Evaluation of Various Machine Learning Classifiers." International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology. Singapore: Springer Nature Singapore, 2022.

[28]. Jahnavi, Y., et al. "Prediction and Evaluation of Cancer Using Machine Learning Techniques." International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology. Singapore: Springer Nature Singapore, 2022.

[29]. Y. Jahnavi, A New Term Weighting Algorithm for Identifying Salient Events (LAP LAMBERT Academic Publishing, 2018)

[30]. Y. Jahnavi, Data Classification using Waikato Environment for Knowledge Analysis (LAP LAMBERT Academic Publishing, 2019)

[31] Jahnavi, Yeturu, et al. "A Novel Processing of Scalable Web Log Data Using Map Reduce Framework." Computer Vision and Robotics: Proceedings of CVR 2022. Singapore: Springer Nature Singapore, 2023. 15-25.