International Journal of Engineering & Technology, 7 (3.24) (2018) 724-732

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comparative analysis of machine learning algorithms applied to multi lingual texts summarization

Archana N.Gulati[1]*, Dr.Sudhir Sawarkar[2]

Department of Computer Engineering, Datta Meghe COE, Navi Mumbai, India
**Corresponding author E-mail: ang.cm.dmce@gmail.com*

Abstract

Over the scarce period the World Wide Web (WWW) takes prolonged extremely and huge volumes of information in the form of news articles is available online. Many a times individuals don't take the spell besides tolerance towards recite whole news divisions or ample long articles. At this time ascends the essential of computerized texts summarization. Uncertainty an instant of the real fillings of the broadcast object is obtainable formerly it will convert calmer for the handler to get a gist of the article as well as it would save a lot of his time. Nearby, numerous methods towards texts summarization which could be off the record on the root of numerous factors such as level of processing, kind of information being processed, etc. The work proposed in this paper tries to integrate these approaches with modern computational linguistics, semantic technologies and machine learning algorithms to devise a novel technique for multi lingual text summarization which could produce summaries aimed at sole too as group of forms. The anticipated method specifically addresses two major languages for the study, one is English being the language used worldwide and second Hindi being the national language of India. The machine learning techniques used for extraction are neural networks and fuzzy logic systems. Finally, a comparison of these techniques is done to show that fuzzy logic systems give better precision as compared to neural networks for summarization in both the languages. The average difference in precision is around 8-10% for Hindi and around 45-50% for English text documents.

Keywords— *multi lingual text summarization, computational linguistics, machine learning techniques, semantic technologies*

1. Introductions

During the development of World Wide Web takes led to group of marvelous amounts of info completed the Internet. Here are everyday e-newspapers which remain obtainable covering newscast on numerous themes similar policy, sporting, confidential etc. Furthermore, informs of specific newscast can seem in the broadsheet for numerous beings and in numerous newspapers through diverse statistics. Then numerous periods persons impartial famine an essence of what the newscast situation is as they don't consume period to recite the whole news trainings owed to absence of period. Now rises the essential of computerized texts summarizations. Such systems generate a condensed form of the exclusive texts which makes the task simpler and easier for readers with lack of time and patience.

Although there are systems which generate summary for multiple documents and for multiple languages, there's no such system developed that can generate efficient summary for multiple document and for languages like Hindi.

Moreover, systems which work with semantic features are also few and have not given considerable results. Considering these parameters, a summarization system that can mull over certain semantic characteristics of the text along with the existing statistical ones and can also handle multiple documents and multiple languages at the same time has been built. A comparison of two different machine knowledge methods i.e., neural's in addition fuzzy's remains practical toward the Hindi and English text to show that fuzzy logic systems give a better precision as compared to neural networks with a difference of about 8-10% for Hindi and around 45-50% for English text documents.

2. Literature Survey

Junlin Zhanq's along with Le Sun's and Quan Zhou's [1] have proposed a Hub's Authority frameworks where the sub_topic in multi_document remain detected through verdict grouping besides the features and words from dissimilar sub_topic is mined. A graph is generated where all the feature words are marked by way of the vertices of the Hub then altogether the judgements remain stared by way of the vertices of Expert witness. Uncertainty a verdict covers the disputes in Hubs, an advantage is shaped among the Hub term and the Expert verdict. By the mutual reinforcement mechanism of the Hub-Authority algorithm, ranking of the sentences is done which is based on cue phrases and sentence length within the multi-documents. Finally, the Markov Representations are cast-off to command the sub_topics within the concluding summary.

Yan-Min Chen along with Xiao-Long Wang and Bing-Quan [2] in their work have used lexical chains for indicative summarization of manifold forms printed in Chinese. The technique calculates verbal chains created proceeding the How_Net information database to progress the presentation aimed on Chinese's languages. Constructed on HowNet knowledge, similarities between sentences and redundancies in text could easily be identified and removed. The algorithms pre_processe the manuscript, before lexica's chain are constructed and finally strong chains are identified. Lastly, the instantaneous remains made in ascending directive, then the anaphora determination knowledge remains practical to improve the articulacy of immediate. Analysis of the method shows that the

lexical chains are effective for generating multiple_documents summary.

Cem Aksoy along with Ahmet-Bugdayci, Tunay-Gur, Ibrahim Uysal and Fazli Can [3], in their paper have proposed a technique to abundant the judgements created arranged the significance of semantics phrase's they encompass. The significance of a phrases remains resolute through its frequencies. The method called Semantics Role Labeling (SRL) tries toward classify the ingredients of a verdict, composed through their characters with deference to the verdict establishes. The experimentation was carried on DUC 2004 dataset and was seen that SRL'_-based sentence counting method outstrips altogether the approaches except the term-based approach. Further it was also found that when SRL's_based scorings are combined through MEAD by way of a auxiliary feature, the routine of the system amplified.

Liang Ma along with the Tingting-He, Fang-Li, Zhuomin-Gui, and Jinguang-Chen [4], propose a method called inquiry absorbed multiple_documents summarization. Here those sentences in the document whose keywords that match the query judgement remain designated for presence in the concluding instant. Two features are calculated one is the query connected features then the additional is the theme linked feature aimed at each term in the article sets. Then the standing of the term is gained by merging the grooves of the two topographies. The concluding notch of each verdict is calculated over the position of arguments which they hold. This method works well for both DUC 2005 and DUC 2006 document corpus.

Yong Liu, Xiao lei Wang, Jin Zhang, Hongbo Xu [5], have devised a method grounded on Modified Pages Ranks. In this technique firstly, a salience model is trained founded on Naïve's Bayes using worldwide features of the sentences like paragraph, position in paragraph, mood and length. Further a relevance model is created where judgements straight applicable to the query are selected. Then using both these models a modified prior possibility aimed at apiece judgement is calculated. Through the assistance of this modified previous probability, a Modified Page_Rank position procedure is agreed out dependent on the relations amongst entirely sentences in the quantity. Likewise, the idleness issue is correctly measured. The closing instantaneous is made by choosing the verdicts with together in height query absorbed info and better data innovation.

Sun Park, ByungRae Cha [6], propose a summarizations technique in which non_negative matrix-factorizations (NMF) then NMF's grouping, is cast-off to abstract expressive verdicts after querys_created, multiple_documents. In these algorithms a verdict is decayed hooked on the linear mixture of non_negative semantics features to mean a sentence's as the totality of individuals semantics features. As a result more meaningful sentences can be extracted which are closest to the query. The advantage of using NMF clustering is it avoids noise and helps to remove redundant sentences. Nevertheless, this method also enhances the accuracy of summarization through categorization mined sentences' in the directive of their abundant.

Hongling Wang Guodong Zhou [7], in their broadside presents a topics_driven outline aimed at generating a multi document instantaneous. Here the assumption made is that the rapid probability dispersal done with the topics should remain reliable with the multiple_documents probability dispersal over the essential topics. The themes remain derived thru Latent-Dirichlet Allocations after multiple_document then is well-defined by way of subjective 'bags_of_words'. Therefore, it becomes easier towards excerpt a verdict before immediate by manipulative the resemblance among a sentence and summaries then the specified multiple_document through their theme probability deliveries. However, measuring the similarity, two methods are used i) the static method which used toward notice the saliences of data in addition ii) the dynamic method to controller idleness in a lively method. Numerous general structures are used toward recover the routine and it is seen that this summarization method works well on TAC.

Harsha Dave and Shree Jaswal [8], introduced a hybrid method for multi document script-summarizations which contains of dual main stages: first is the extractive summary besides second is the abstractive one. In the extractive phase firstly, linguistic analysis is carried out, then preprocessing and removal of redundant sentences and finally summary generation by extracting those sentences whose term frequency score is above a predefined threshold. Later on this extracted summary is applied the abstractive summarization process. It involves generating a word graph based on domain ontology and WorldNet and applying certain heuristic rules to generate a final reduced summary. The abstractive summary generated is understandable and readable. The only drawback being that as the size of the document increases the time required to generate the summary increases.

Lei Yu, Fuji Ren [9], have applied four models to Chinese and Japanese text. The four trainable models used were Conditional_Random_Field (CRF) then Hidden-Markov Models (HMM), Mathematical-Methods of Statistics (MMS) and Gaussian Mixture Models (GMM). For training the model, several structures such for example sentence location, sentence uniqueness, and quantity of Name Object etc. remain cast-off. Dissimilar firmness rates such as 10%, 20%, 30% were used to measure the precision. It was observed that HMM/CRF/GMM display healthier consequences than MMS taking place mutually Chinese then Japanese writing. Furthermost highly, GMM's provided the finest accuracy.

Gael-de-Chalendar along with the Romaric_Besan, Olivier-Ferret, Gregory Grefenstette then Olivier Mesnard [10], as per today's need since the WWW is becoming widely multilingual, the authors have introduced a system for generating cross lingual summaries .To achieve this the documents in their original language undergo a thematic analysis treatment which is based on the most frequently used thematic words in the text document. These sentences extracted further underwent a syntactic analysis for further simplification which is represented using dependency relation graphs. From these dependencies, finally a text output is generated in another language. The reordering of text in the final summary involves employing precedence rules learned from independent source language text.

Ha-Nguyen_Thi-Thu along with-Quynh-Nguyen-Huu, Tu- Nguyen-Thi-Ngoc [11], have proposed an extractive text summarization technique for Vietnamese language. They have used neural network for supervised learning to decide upon which sentences should be included in summary and which shouldn't be included. To overcome the cost and reduce the computational complexity they have also combined dimensional feature reduction when building term sets. Since not much work has been done on Vietnamese language, hence standard corpus was not available and therefore manually generated documents were used for training and testing purpose.

Jayashree R, Srikanta Murthy K , Srikanta Murthy K [12] suggest a neural network based summarizer for text documents in Kannada .Error back propogation algorithm is used for training using Kannada text documents which have been custom built. Human readers give the judgement on whether a sentence should remain encompassed in instantaneous otherwise not. That stands the biggest advantage of the method that the neural network can be trained according to the reader's style.

Sakshee Vijay, Vartika Rai [13] have projected the extractive summarizations technique. For choice of significant verdicts after the unique text, where word level then sentences level topographies remained appraised. According to them proper selection of features and heuristics leads to better summarization and the same has been achieved in their work.

Manisha Gupta and Dr.Naresh Kumar [14] projected a method to produce extractive instant aimed at only Hindi text documents by means of a rule based method along with dead phrase and deadwood removal. 96% accuracy and 60-70 % compression is achieved.

K.Vimal Kumar along with the Divakar Yadav then Arun Sharma [15], have proposed an extractive text summarization approach for Hindi language text documents. The basic idea of this method is not just to extract sentences but also to check their relevance with other sentences semantically. So that if the sentences are semantically similar they can be replaced by just one sentence rather than repetition. For this purpose a graph based approach is used to link the sentences and generate the summary in a meaningful way. The system achieved a precision of 79% and recall of 69% which is reasonably good proportion.

Luhn's method [16] focuses on frequency and position of the most significant words. The method is economical, simple and easy to implement but the time complexity is very high. Luhn proposed this method for limited capability machines and hence semantic properties of text were not considered .His emphasis was only on frequent words. According to him, minimum and maximum threshold for frequency of words can be set and a comparison can be done with the common wordlist. This method was also proposed for only single English text documents.

Udo Hahn, Inderjeet Mani [17] have come up with the challenges of automatic summarization. They have considered both the methods of summarization i.e. extractive and abstractive and discussed the challenges like the reduction rate, evaluation criteria for the summaries generated. According to the authors the methods of creating and evaluating a summary must complement each other. Secondly they say a lot of background knowledge is needed to achieve high reduction. Most of the methods apply linear weighting models which weight individual sentences for different features and then find the overall weights by summing up the individual weights. These weights will help in determining that by which sentences to contain in the summarys.

Ladda-Suanmali along with the Mohammed Salem, -Binwahlan and Naomie-Salim [18] proposed a method for extractive summarization of a document by using important features based on fuzzy logic. A comparison with MSWord and baseline summarizer showed that the proposed method worked far better for single document text summarization.

Jyoti Yadav and Yogesh Kumar Meena [19] have proposed a methodology for improving excellence of instantaneous by means of fuzzy logic, bushy technique and WorldNet synonym's towards grip the matters of ambiguities and inexact standards. A comparison of this new method with the individual methods showed considerable improvement in summarization for single document.

S. Santhana Megala along with the Dr. A. Kavitha and Dr. A. Marimuthu [20] uses the abstraction procedure aimed at verdict choice. At this time roughly feature built sentence keep count methods are similarly used, which frolicked an significant character in the manuscript summarizations. Lastly an examination is complete by associating the fuzzy_logic formerly neural network methods built upon the exactness, memory and F_measures. Fuzzy_logic directions remained cast-off to steadiness the bulks amongst significant then insignificant features built on the features-extraction.

3. Proposed Methodology

A summarization system that can handle multiple documents and multiple languages at the same time has been built. The languages considered are English and Hindi because English is used worldwide and Hindi is the national language of India. This system considers a combination of statistical as well as linguistic and semantic features of the text to generate more accurate summaries. Now, fuzzy-logic then neural-networks remain functional separately on that the text and then a comparison of these methods is done to check which method gives better summarization. Thus, an original framework aimed at multiple_document, multi_language then extractive summarizations remains developed. The basic flow of the system is as shown in fig.1.The proposed system has been implemented in two parts:

1. Text summarization using fuzzy logic.
2. Text summarization using neural networks.

A. Proposed Method 1: Text summarization using hybrid features and applying fuzzy logic to English as well as Hindi dataset.

Appropriate towards instrument the texts summarizations created on fuzzy_logic method, Java implementation software package of fuzzy control systems called jFuzzylogic was used. The aforementioned gears comprehensive fuzzy interpretation scheme too by way of fuzzy switch sense [9]. The algorithm works as follows:

Step 1: Initially the documents are preprocessed to bring them into a form that makes the task of summarization easy.

Step 2: Next, the hybrid features and their score is extracted and remain cast-off as say towards the fuzzifier.

Step 3: An input triangular membership function is used and for each feature five fuzzy sets are defined which are composed of insignificant value (low (L) besides identical low (VL) amd the Median (M) / significant values (high (H) then same high (VH).

Step 4: A value in the range of 0 to 1 is gained aimed at each sentence in the productivity, founded scheduled verdict landscapes besides the obtainable fuzzy instructions clear in the information sordid.

Step 5: The previous stage in fuzzy_logic classification remains the defuzzifications. The centroid based output membership function is used for the purpose. This remains separated hooked on 3 association function:- Output a.Unimportant/ b.Average/ c. Important to change the fuzzy consequences from the inference engine into a crisp output for the final score of each sentence.

Step 6: The gained price in the production controls the grade of importance of the sentence and thus the inclusion/rejection of the sentence in the final summary.

Step 4: A value in the range of 0 to 1 is gained aimed at each sentence in the productivity, founded scheduled verdict landscapes besides the obtainable fuzzy instructions clear in the information sordid.

Step 5: The previous stage in fuzzy_logic classification remains the defuzzifications. The centroid based output membership function is used for the purpose. This remains separated hooked on 3 association function:- Output a.Unimportant/ b.Average/ c. Important to change the fuzzy consequences from the inference engine into a crisp output for the final score of each sentence.

Step 6: The gained price in the production controls the grade of importance of the sentence and thus the inclusion/rejection of the sentence in the final summary.

B. Proposed Method 2: Text summarization using hybrid features and applying neural network to English as well as Hindi datasets.

Encog, a machine learning framework for Java has been used to support the neural networks. The neural network algorithm[12] for summarization works as follows:

Step 1: Initially the documents are preprocessed to bring them into a form that makes the task of summarization easy.

Step 2: Next, the hybrid features and their score is extracted and are used as input to the neural network.

Step 3: The training process involves using the back propagation network to learn the types of sentences that should be included in summary. Back propagation is one of the oldest training methods for feed forward neural networks and gives reasonably good results. This is done by training network with sentences from several text documents. With the help of input neurons, calculated value of features for sentences is fed to network.

Step 4: Further calculation in the hidden layer is performed by applying the activation function and error calculation and updation of weights. This procedure is repeated for multiple epochs till the error achieved is 0.01. During every cycle updation of weights takes place.

Step 5: Finally, the output neuron defines one of the two output classes either 0 or 1 indicating whether the verdict must be comprised in swift otherwise not.

Step 6: All those sentences marked as 1 are selected for inclusion in the final summary.

Around 50 news articles on sports and politics were cast-off towards training the neural network and an additional group of 20 documents

were taken for testing

C. Hybrid features set used for text summarization purpose.

F1: TFe_Inv.SentFe (Term_Frequency_Inverse Sentence Frequency): TFe_Inv.SentFe is calculated to judge the position of a sentence by finding out the rate of recurrence of a word within a sentence and it is calculated as

$$TFe\_Inv.SenFe = TFe \times Inv.SentFe \quad (1.a)$$

$$TFe = \frac{F}{N} \quad (1.b)$$

F equal the frequencies of the words in the sentence

N=No. of words in that sentence

$$Inv.SentFe = log[\frac{Ns}{SF}] \quad (1.c)$$

Ns == Total number of the sentences in a document

SF== Sentence of frequencies, remains count of the sentence in which the words has happened in a documents of N sentence.

Average TeFe is considered for each sentence then accordingly the weight score is allocated to the aforementioned.

F2:-Length of sentence in the documents (SeLen):

Usually, very short or very long sentence do not own any remarkable info then remain not beneficial in instant group. These features are glowing clarified through the formulation

$$SeLen = sin\{(L - MinL) \times [\frac{(Max\theta - Min\theta)}{(Max\ L - Min\ L)}]\} \quad (2)$$

Otherwise

$SeLen=0 \quad if \quad L < MinL \quad or\ L > MaxL$

Where, L = Length of Sentence

Empirical observations show that Minimum length (MinL) should be atleast three words so as to form the shortest sentence making sense whereas and maximum length (MaxL) can be reserved fifteen words since after this edge, there are more chances of redundancy.

θ ranges from zero to 180 deg. to become suitable outcomes.

F3: Location of sentence in the document (SLoc):

The position of the sentence in the document is also an important parameter in deciding the inclusion of sentence in summary. Sentences in the starting of the text outline the resolution of the text while final sentences typically take an essence of the text. The edge value chooses in what way numerous verdicts after the start and finish must be reserved in the instant. The features groove can be glowing distinct through

$$SLoc = cos\{(L - MinV) \times [(\frac{Max\ \theta - Min\ \theta}{Max\ V - Min\ V})]\} \quad (3)$$

Else,

SLoc = 1(if within threshold range)

Where

MinV =Minimum no. of sentences= Ns * TH

MaxV=Maximum no. of sentences = Ns * (1 - TH)

Ns == Total no. of sentences with document

TH == Threshold_Value

Min θ == Minimum Angle == Zero degrees

Max θ == Maximum Angle ==180 degree

CL == Current_Locations of sentence

F4: Word Similarity between sentences (Sim (Si, Sj)):

Word Comparison among 2 sentences remains found through stemmed word similar. The aforementioned is assumed by the formula

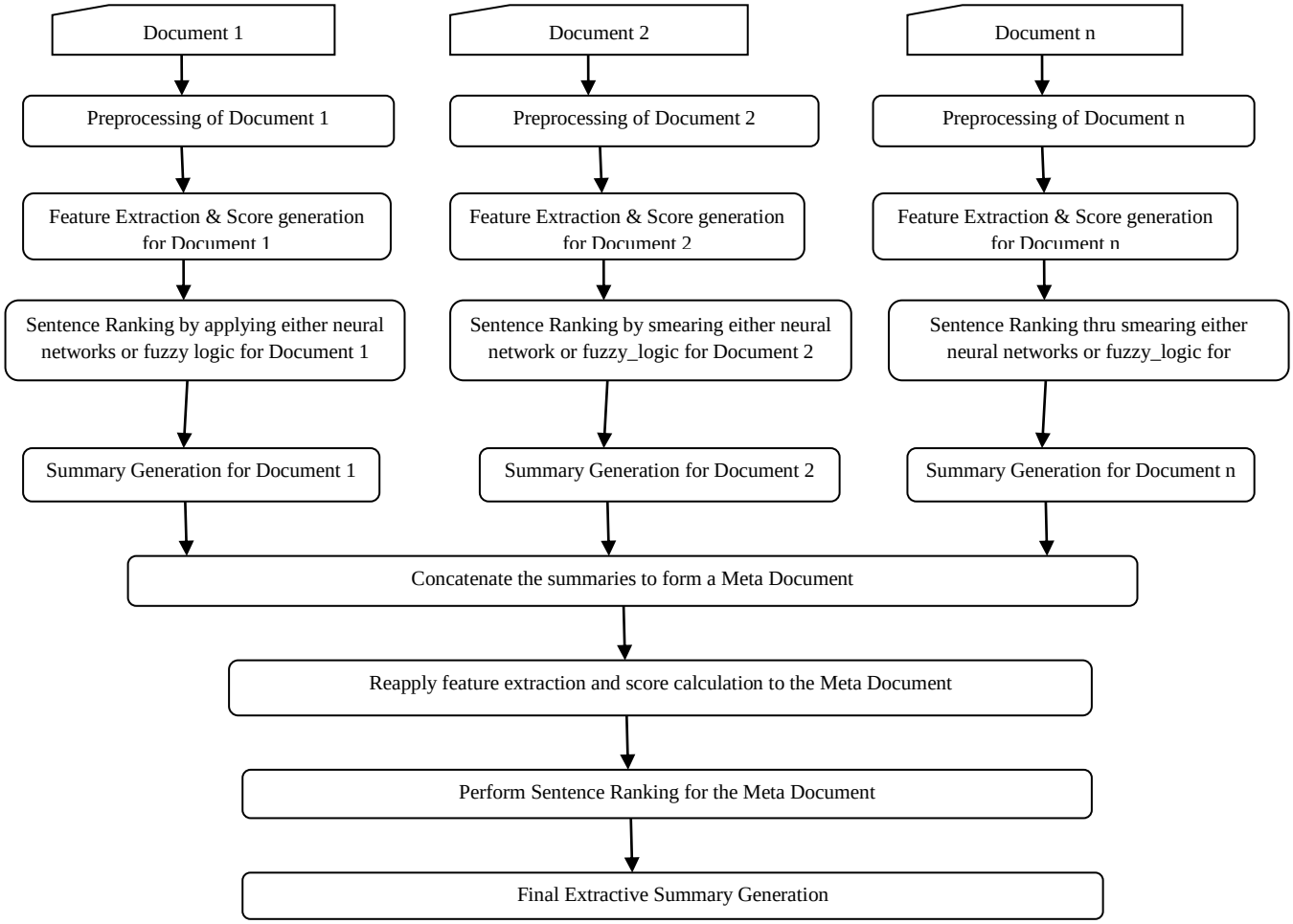

Fig. 1. Generalized System Flowchart

$$Sim(Si,Sj) = \frac{\frac{No.of\ words\ occured\ in\ sentence\ Sj\ that\ exist\ in\ Si}{Wt}}{} \quad (4)$$

where,

Wt == Total number of word in a sentence Si

The higher the score, more is the similarity between sentences.

F5: Numerical Data (Num)

Sentences containing numerical data more often convey some important information like important dates, figures, denomination, scores and they must be comprised with the summary. The Weightiness aimed at this feature is intended as

$$Num = \frac{No.of\ numerical\ data\ tokens\ present\ in\ Si}{Total\ no.of\ words\ in\ the\ sentence} \quad (5)$$

F6:- Title_Overlap (TO(Si))

Title generally gives us a depiction of what might be the gratified of the documents. So, uncertainty a sentences Si consumes better contest through the title confrontations then we roughly Si remains more significant than other verdicts in that text.

$$TO(Si) = \frac{Nt}{Wt} \quad (6)$$

Nt == No. of word of Si happened in heading

Wt == Total Number of words within Sentence Si

F7:- Subject_Object_Verb (SOV) and Subject_Verb_Object (SVO) Qualifier

Hindi is named as SOV linguistic for the reason that maximum of the sentences takes the method <subject><object><verb> while in English the sentences procedure remains SVO <subject><verb><object>. Aimed at a sentence designate SOV or else SVO qualified, individually word in the sentence consumes designate marked by conveying suitable fragment of language (POS) similar (Noun/Adjective/Verb/Adverb). Aimed at POS classification Essential NLP tool established thru Stanford has stood cast-off. Built on the labels allocated, the initial noun word in the sentences remains noticeable by way of theme of the sentence. The complete sentence remains analysed till the termination, till an object_verb or else verb _object couple is found, besides if originate the sentence remains supposed designate SOV and SVO competent. Advanced the score of the sentence, more important it is for summary.

SOV(Si) or SVO(Si) = 1(if SOV/SVO qualified)

Else

$$SOV(Si)\ or\ SVO(Si) = O \quad (7)$$

F8:- Subject_Similarity

In the subject resemblance features, theme of the name remains likened through the subject in the sentence. If a subject match found, the sentence is given more credence.

Sub_(Si) = 1, if POS is the noun/root value of title and sentence are equal

= 0, otherwise (8)

F9:- Cue_Phrase

Cue_Phrase remain that specific keyword which remain actual cooperative in determining sentence position. A tilt of about 100's Hindi then English prompt phrases partakes stood mined in addition individual sentences covering these signal expressions are measured additional significant besides assumed an advanced weightage.

$$Cue(S_i) = \frac{No.of\ cue\ phrases\ present\ in\ the\ sentence}{Total\ no.of\ words\ in\ the\ sentence} \quad (9)$$

F10: Thematic words

Other than the stop words, there are words which are more frequent in the text document and they are more relevant to the topic or theme of the document. Such top 15 frequent words are identified and sentences having these words are given a higher score.

$$Them(Si) = \frac{No.of\ thematic\ words\ in\ the\ sentence}{Total\ no.of\ words\ in\ the\ sentence} \quad (10)$$

F11:- Lexicals_Similarity

Assumed 2 sentences, the lexical similarity controls in what way comparable the connotation of 2 sentences remains. The advanced the notch, the additional alike the sense of the two sentences remains.

The similarity measure between two words w1 and w2 is calculated as

$$LS(w1,w2) = \frac{1}{d\_lcs(w1,w2)} \times \frac{1}{2}\left[\frac{d(w1)-lcs(w1,w2)}{path(w1)} + \frac{d(w2)-lcs(w1,w2)}{path(w2)}\right] \quad (11.a)$$

where,

d_lcs is the depth of lowest common subsummer from root node.

D is the depth of the term from root node.

path is the no. of paths going through the word w towards the root node.

Further, the lexical similarity between two sentences is calculated as,

$$SS(S_x,S_y) = \frac{\sum_{x=1,y=1}^{x=N,y=N}\left[\sum_{j=1}^{m} LS(Sxw1,Sywj)\times\sum_{j=1}^{m} LS(Sxw2,Sywj)\times....\times\sum_{j=1}^{m} LS(Sxw\right.}{N} \quad (11.b)$$

where,

S_x and S_y are any two sentences in the document.

S_xw_1, S_xw_2,..., S_xw_i are words in sentence S_x

S_yw_j are words in sentence S_y varies from 1 to m

N remains the total number of sentences in the text

4. Experiment Atal Results

For experimentation purpose datasets used were Hindi and English news articles from online sources such as Google news. Final evaluation of system generated summary with human generated summary in terms of Precision/Recall then F-Score and remain distinct as,

i)Precisions

Precisions remains the no. of precise sentences alienated through no. of verdicts mined.

ii)Recalls

Recalls are no. of precise sentences divided thru the no. of sentences that must consume remained mined.

iii)F_scores

F_scores remains the vocal callous of precision and recall.

Table I, Figure.2 and Figure.3 demonstrations the contrast of evaluation results for various datasets in the method in the precisions/recalls then f_score aimed at Hindi scripts forms smearing fuzzy logic and neural networks.

Table I: Comparison Of Fuzzy Logics & Neural Network For Hindi Text Documents

	Fuzzy			Neural		
	Precisions	Recalls	F_Scores	Precisions	Recalls	F_Scores
Dataset 1	62.50	71.43	66.67	60.00	42.86	50.00
Dataset 2	66.67	56.66	61.25	57.56	45.50	51.31
Dataset 3	75.00	69.66	72.23	74.54	68.96	71.64
Dataset 4	50.00	66.67	57.14	32.50	55.60	44.05
Dataset 5	72.30	68.50	69.67	66.26	59.56	62.91
Average	65.29	66.58	65.39	56.17	53.42	54.55

Table Ii. Comparison Of Fuzzy Logics & Neural Network For English Text Documents

	Fuzzy			Neural		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Dataset 1	75.00	66.67	70.59	25.00	11.11	15.38
Dataset 2	54.55	66.67	60.00	38.46	55.56	45.45
Dataset 3	64.45	68.00	66.15	29.24	20.22	23.91
Dataset 4	80.55	70.64	75.27	48.36	55.56	51.71
Dataset 5	69.29	67.76	68.51	41.35	62.16	49.66
Average	68.77	67.95	68.10	36.48	40.92	37.22

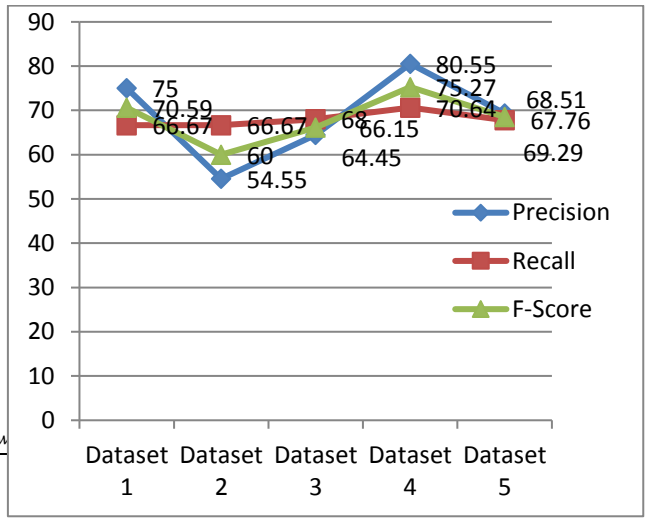

Fig. 2. Graph representing values of precision, recall and f-score for Hindi documents using fuzzy logic

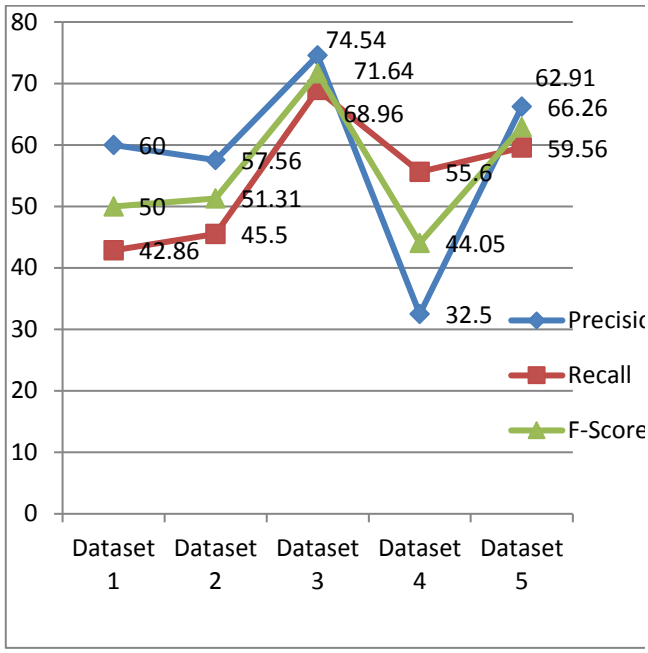

Fig. 3. Graph representing values of precision, recall and f-score for Hindi documents using neural networks

Fig 4 represents the comparison of average precision, recall and f-score values for Hindi documents using fuzzy logic and neural networks over five sample datasets.

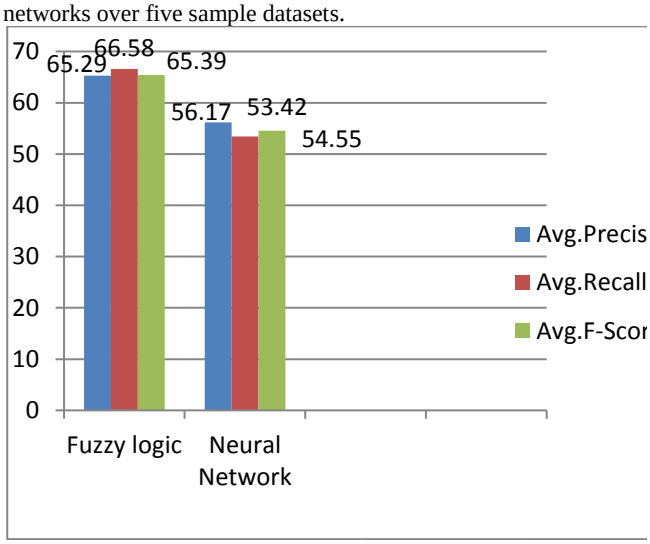

Fig. 4 Graph representing comparison of average values of precisions, recall and f-score for five sample Hindi datasets using fuzzy logic and neural networks.

Table II, Fig 5 and Fig 6 show the comparison of evaluation of results for various datasets in the form of precision, recall and f-score for English text documents for fuzzy logic and neural networks.

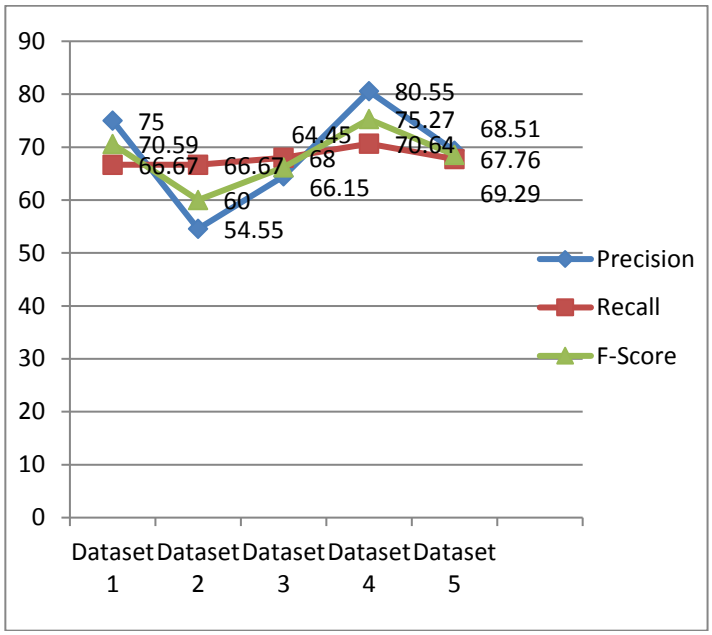

Fig. 5. Graph representing values of precision, recall and f-score for English documents using fuzzy logic

Fig 7 represents the comparison of average precisions, recalls and f-scores values aimed at English documents using fuzzy logic and neural networks over five sample datasets.

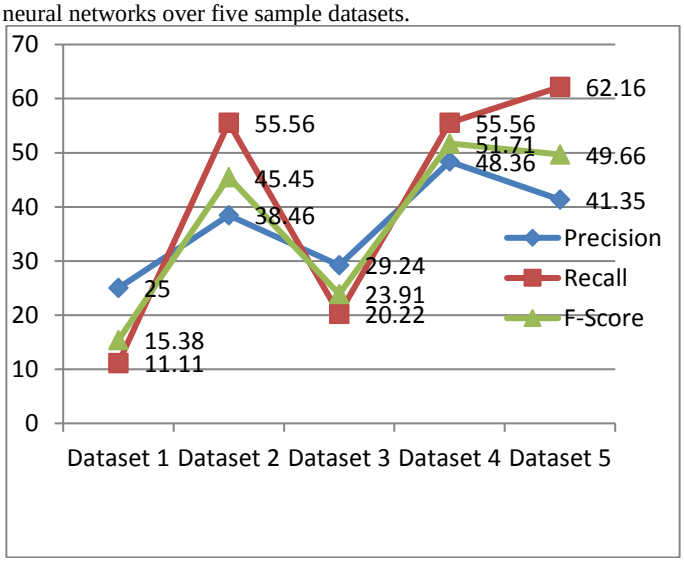

Fig. 6. Graph representing values of precision, recall and f-score for English documents using neural networks

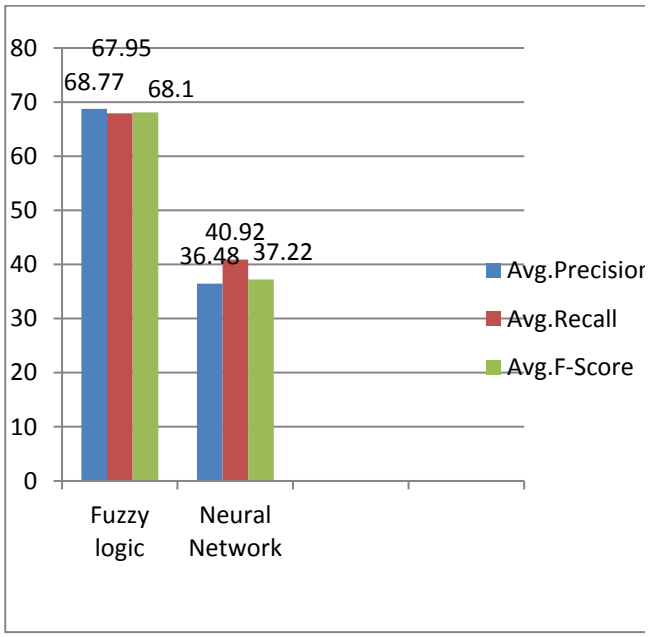

Fig.7. Graph representing comparison of average values of precision, recall and f-score for five sample English datasets using fuzzy logic and neural networks.

Also, the machine learning algorithms were applied on various Hindi and English text documents with and without considering the lexical features(F9,F10,F11) . which showed that adding lexical features gave better precision.

Table III shows the comparative analysis of average values of precision, recall and f-score for eight and eleven feature set using the two different machine learning algorithms.

Table Iii. . Comparison Of Fuzzy Logic And Neural Networks For English Text Documents

	Basic feature set			Basic +lexical feature set		
Average values	Precision	Recall	F-Score	Precision	Recall	F-Score
Hindi-fuzzy	65.29	66.58	65.39	68.64	63.80	65.92
Hindi-neural	56.17	53.42	54.55	60.56	56.37	58.39
English-fuzzy	68.77	67.95	68.10	74.64	70.00	72.25
English-neural	36.48	40.92	37.22	44.36	48.54	46.36

The present work is compared with existing methods for single document English text summarization [18], [19], [20] as shown in Table 4 and Hindi text summarization [14], [15] as shown in Table 5 .It shows that proposed method 1 gives improved average precision as compared to almost all other methods whereas method 2 fails to achieve the desired accuracy.

Table 4. Comparison of average precision for English text documents

Documents	Proposed (1)	Proposed (2)	[18]	[19]	[20]
English text	74.64	44.36	47.589	43.411	48.629

Table 5. Comparison of average precision for Hindi text documents

Documents	Proposed (1)	Proposed (2)	[14]	[15]
Hindi text	68.64	60.56	96.00	79.00

5. CONCLUSION

Thus, two techniques namely neural networks and fuzzy logic were applied over multiple documents to generate extractive summary by using eight hybrid feature set. Appropriate selection of features led to summaries produced through the organization being identical close to summaries made by persons. Moreover, it was also found that fuzzy logic systems work better than neural aimed at together English then Hindi, multiple_document texts summarizations. The improvement in precision varies from 8-10% for Hindi and around 45-50% for English text documents.

Acknowledgment

I express my sincere gratitude to Prof.Dr.S.D.Sawarkar, my guide meant for creation of this effort a achievement. Without his support and guidance, it would not have remained conceivable to comprehensive this task.

References

[1] Junlin Zhanq ,Le Sun, Quan Zhou, A Cue-based Hub-Authority Approach for Multi Document Summarization, Proceeding of NLP-KE'05

[2]Yan-Min Chen, Xiao-Long Wang,Bing-Quan , Multi-document Summarization based on Lexical Chains, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18- 21 August 2005

[3]Cem Aksoy, Ahmet Bugdayci, Tunay Gur, Ibrahim Uysal, Fazli Can, Semantic Argument Frequency-Based Multi-Document Summarization, ISCIS, September 14-16, 2009, METU, Northern Cyprus Campus.

[4] Liang Ma, Tingting He, Fang Li, Zhuomin Gui, Jinguang Chen, Query-focused Multi-document Summarization Using Keyword Extraction, 2008 International Conference on Computer Science and Software Engineering.

[5] Yong Liu, Xiao lei Wang, Jin Zhang, Hongbo Xu, Personalized PageRank based Multi-document Summarization, IEEE International Workshop on Semantic Computing and Systems, 2008 IEEE.

[6] Sun Park, ByungRae Cha, Query-based Multi-document Summarization using Non-negative Semantic Feature and NMF Clustering, Fourth International Conference on Networked Computing and Advanced Information Management, 2008 IEEE.

[7]Hongling Wang Guodong Zhou, Topic-driven Multi-Document Summarization, 2010 International Conference on Asian Language Processing

[8] Harsha Dave and Shree Jaswal, Graph Based Technique for Hindi Text Summarization, 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015),Dehradun, India, 4-5 September 2015.

[9] Lei Yu, Fuji Ren, A Study on Cross-Language Text Summarization Using Supervised Methods.

[10] Gael de Chalendar, Romaric Besan, Olivier Ferret , Gregory Grefenstette and Olivier Mesnard, "Crosslingual summarization with thematic extraction, syntactic sentence simplification, and bilingual generation", CEA-LIST LIC2M, BP6 F92265 Fontenay-aux-Roses France.

[11] Ha Nguyen Thi Thu, Quynh Nguyen Huu, Tu Nguyen Thi Ngoc, A Supervised Learning Method Combine with Dimensionality Reduction in Vietnamese Text Summarization, ©2013 IEEE.

[12] Jayashree R, Srikanta Murthy K , Basavaraj.S.Anami , Suitability of Artificial Neural Network to Text Document Summarization in the Indian Language-Kannada, International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 6 (2014) pp. 626-634 © MIR Labs.

[13] Sakshee Vijay, Vartika Rai, "Extractive Text Summarization in Hindi",2017 International Conference on Asian language Processing.,pp318-321,2017 IEEE.

[14] Manisha Gupta and Dr.Naresh Kumar, "Text Summarization of Hindi Documents using Rule Based Approach",2016 International Conference on Micro-Electronics and Telecommunication Engineering.,pp336-370,2016 IEEE.

[15] K. Vimal Kumar, Divakar Yadav and Arun Sharma," Graph Based Technique for Hindi Text Summarization", © Springer India 2015 ,J.K. Mandal et al. (eds.), Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, pg 339.
[16] Luhn, H. P. 1958, "The Automatic Creation of Literature Abstracts", IBM Journal, pp. 159-165.
[17] Udo Hahn Albert Ludwigs University Inderjeet Mani Mitre Corp.," The Challenges of Automatic Summarization", 2000 IEEE, November 2000.
[14] Ladda Suanmali, Mohammed Salem, Binwahlan,Naomie Salim, "Sentence Features fusion for text summarization using fuzzy logic",2009 Ninth International Conference on Hybrid Intelligent Systems.,pp142-146,2009 IEEE.
[17] Jyoti Yadav and Yogesh Kumar Meena," Use of Fuzzy Logic and WordNet for Improving Performance of Extractive Automatic Text Summarization",2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI),Sept 21-24,Jaipur,India pp2071-2077, 2016 IEEE .
[18] S. Santhana Megala, A. Kavitha, A. Marimuthu , "Enriching Text Summarization using Fuzzy Logic", International Journal of Computer Science and Information Technologies, Volume 5, Issue 1, 2014.