Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

Hybrid Algorithm to Generate Summary of Documents by Extracting Keywords

[1]Surbhi Patel,[2]Devanshi Parikh, [3]Dr. Hiren Joshi
[1]M.Tech Student,[2] M.Tech Student, [3] M.Tech Professor
[1,2 ,3]Department of Computer Science
[1]Rollwala Computer Center, Gujarat University, Ahmedabad, India

Abstract : **In modern times, data is growing rapidly in every domain such as news, social media, banking, education, etc. Due to the excessiveness of data, there is a need of automatic summarizer which will be capable to summarize the data especially textual data from original document without losing any critical purposes, and "Keywords" in a document represents subset of words or phrases from the document for describing its meaning. Manual assignment of quality keywords is time-consuming and expensive. In this paper, we present our preliminary development including sentence similarity index with cosine to measure connected within clusters where keywords. Using such techniques the novel approach strengthens its process and finds hybrid approach to perform the summarization of document along with the keywords identified by the type of document using text mining techniques. Since text summarization process is highly depend on keyword extraction, the overall results are found promising.**

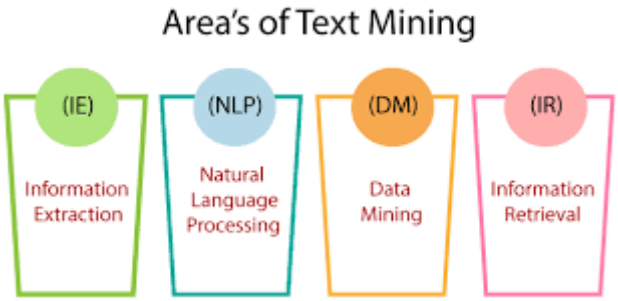
Index Terms– Text Extraction, WordNet, NLP, ML, Sentence Similarity

I. INTRODUCTION

Text mining is the way toward breaking down unstructured text, separating important data and changing it into valuable business intelligence. There is a requirement for a computerized automated framework that can remove just significant data from these information sources. To accomplish such tasks, we have to mine the content from the reports. Data mining and text mining is the way toward removing huge amounts of text or data to determine great values that can help in decision making as well as text filtration for a specific requirements. text mining sends a part of the procedures of natural language processing (NLP, for example, part of speech (POS) labeling, parsing, N-grams, tokenization, and so forth., to play out the content investigation. It incorporates assignments like programmed watchword extraction and content outline. Better methodology is yet to be discovered to analyze valuable text and extract meaning from it. Text mining (TM) used to remove helpful data from an accumulation of records. The way toward examining content to separate data that is valuable for a particular reason. Text mining is like information mining, then again, actually information mining devices are intended to deal with organized information from databases, yet message mining can likewise work with unstructured or semi-organized informational collections, for example, messages, content reports and HTML records etc. Text mining has center around "text".[1]

AREAS OF TEXT MINING

1. Information Retrieval (IR)
    a. Help in deciding limit of the arrangement of records that are applicable to a specific issue.
    b. Accelerate the examination.



TEXT MINING PROCESS

a. Text Identification
b. Text Categorization
c. Text Clustering
d. Text Filtering
e. Text Analyser
f. Predictive modeling

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

1. Text Pre-processing
   i. Text Cleanup Text Cleanup means removing any unnecessary or unwanted information.
   ii. Tokenization Splitting the text.
   iii. Part of Speech Tagging Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text.
2. Text Transformation

A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:

i. Bag of words ii. Vector Space

3. Feature Selection (Attribute Selection)

The process of selecting a subset of important features for use in model creation. Irrelevant features do not provide relevant or useful information in any context.

4. Data Mining

The Text mining process merges with the traditional process. Classic Data Mining techniques are used in the structured database.

5. Evaluate the result.

6. Applications Text Mining can be applied in a variety of areas.

Web mining, Medical, Resume Filtering etc.

In the literature, the key step of this method is to determine which sentences are important to the document, which is usually termed as sentence scoring. Over the years, the field has seen advances in the sophistication of language processing and machine learning techniques for sentence scoring [Nenkovaetal. (2011)]. In the extractive summarization, the summarizer takes input as text file and tokenization of an input text is done in-order to remove find the terms of the text. Then stop words are removed in order to filter the text. And finally, part-of speech tag is added to each token. Abstractive text summarization can solve this problem by representing the extracted sentences into another more understandable semantic form[2].

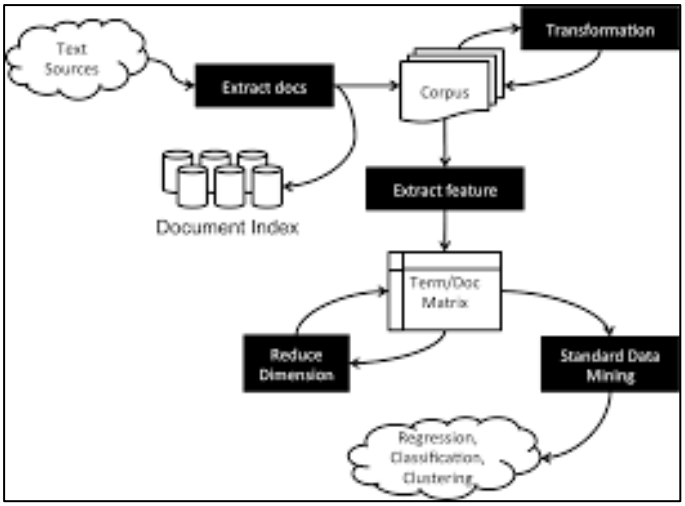
Single document text summarization is to build summary from single source document. This type of text summarization technique accepts only one document as input, then uses different techniques to extract important sentences from source document and then after from extracted sentences summary to be generated. Generation of summary is in more understandable, syntactically or semantically correct and most important in reduced form.

The rest of the paper is organized as follows: In Section II, the motivation of this work is presented. In Section III, the related research in recent years on the recommendation algorithms is introduced. The explanation of terminologies, methodologies and proposed algorithm methods are presented and analyzed in Section IV. Section V presents algorithm of our proposed method. In Section VI, we verify result with experiments. Finally, the conclusion of the paper is shown in Section VII.

II. MOTIVATION

Main aim of any research is to generate end results which are efficient than the earlier one. Manual assignment of quality keywords is error-prone, time-consuming and expensive. While extracting keywords , the methods and algorithms may perform differently. Automatic keyword extraction enables us to identify a small set of words, key phrases, keywords, or key segments from a textual document with the help of which the meaning of the document can be described. Outline of documentation is a productive and ground-breaking strategy that give the short summary after effect of the entire information. In earlier researches, using grouping of multiple methods such as, automatic keywords extractors: Text Rank, RAKE, TAKE are Applied, and calculated parameters: extracted words, correct keywords, precision, recall and f-measure. It has the highest recall and highest fmeasure compared to any of the individual automatic keyword extractors. I found that Precision has lower recall [5]. For Supporting data driven access, C value method ,FGB algorithm approach of these algorithm can be explained more clearly. Some ranking methodology can be helpful to yield better results [6]. RAKE, Text Rank only consider a single document at a time when extracting terms instead of the entire corpus. Results affected by four major types of errors, Over generation errors, Redundancy Errors, Infrequency errors, Evaluation Errors [8]. KBRS - Keyword Based Recommendation System in Social Networks-applied KBRS algorithm (using Hadoop) and UCF Algorithm. Pictures, images and words in such format can be identified and processed [9].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020



III. RELEVANT WORK

In the previous papers as mentioned in the list of references, the text mining process has been analyzed in different ways. A novel statistical method to perform an extractive text summarization on single document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented[1]. In second paper the techniques are based on natural language processing, data mining and semantic similarity domains. These all techniques are used for to generate summary automatically from source document. In their proposed approach provides automatic feature based extractive heading wise text summarizer to improve the coherence thereby improving the understandability of the summary text. Here they have used intrinsic evaluation. The recent approaches are recent literature on automatic keyword extraction and text summarization are presented since text summarization process is highly depend on keyword extraction. This literature includes the discussion about different methodology used for keyword extraction and text summarization. It also discusses about different databases used for text summarization in several domains along with evaluation matrices. In some word sentence co-ranking model, researcher address the sentence scoring technique, a key step of the extractive summarization. Specifically, we propose a novel word sentence co-ranking model named Co-Rank, which combines the word-sentence relationship with the graph-based unsupervised ranking model.

Some extractive summarization techniques are combining the topic of text mining and its relationship with text summarization. important parameters for extracting information from different subject topics are analyzed with different approaches. Another approach is to find 5 different verbal agitation with various text mining techniques and combining acoustic signal processing with three different text mining paradigms. WordNet is the most used and popular dictionary for semantic meanings of words predominant sentences, identified the main stages of the summarizing process, and the most significant extraction criteria are presented. Document corpus are DUC dataset with its wide range of docs and other documents which can be of any valid data sets.

Pattern identification for a document is a analyzing the type of data along with its patterns for its numeric and categorical nature for statistical calculation. Kaggle, github and other google verified sources are a good resources for dataset for testing purpose. It's useful to test the algorithm while retrieving meaningful data from unstructured data.

This paper proposed a methodology where keywords and synopsis of subset of archive could be consequently created during an quick details of meetings or minutes to encourage client's interest looking for process. Right now, the data presents in our early improvement including another techniques for highlighted words extraction and synopsis at the same time over a subset of archives and visual portrayals of those outcomes to help client investigations [5]. Another method proposed for the automated keyword extraction framework is Thai dictionary based arrangement framework which can consequently refresh the word reference and classify them in Thai. The word reference is an collection of vector which is made from the programmed keyword extraction framework.

preliminaries

This section explains the methods required in our work. In our work we are taking into consideration the documents of different types as dataset. We are considering different docuemnts as well as words, phrases and sentences related to one topic.

A. N-Gram

N-gram tokenization process converts words into tokens for further processing. N-grams are used for a variety of different task. For example, when developing a language model, n-grams are used to develop not just unigram models but also bigram and trigram models. Google and Microsoft have developed web scale n-gram models that can be used in a variety of tasks such as spelling correction, word breaking and text summarization. N-grams of writings are widely utilized in content mining and regular language handling assignments. They are essentially a lot of co-happening words inside a given window and when figuring the n-grams you normally push single word ahead

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

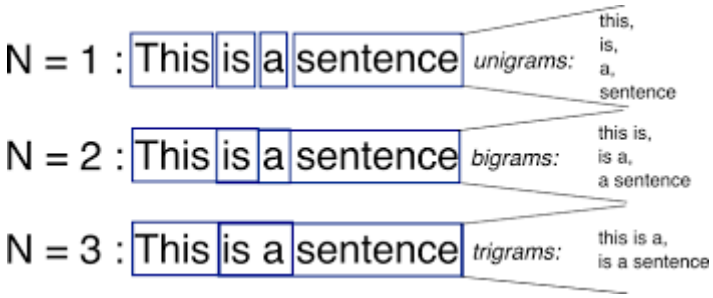


Fig. 1 The n-gram techniques

Let's assume a bigram model. So we are going to find the probability of a word based only on its previous word. In general, we can say that this probability is (the number of times the previous word 'wp' occurs before the word 'wn') / (the total number of times the previous word 'wp' occurs in the corpus) =

```
(Count (wp wn))/(Count (wp))
```

Let's calculate the probability of the word "Diego" coming after "San". We want to find the P (Diego | San). This means that we are trying to find the probability that the next word will be "Diego" given the word "San". We can do this by:

```
=(No of times "San Diego" occurs) / (No. of times "San" occurs)

= 2/3

= 0.67
```

Example of bi-gram

B. Stopping and stemming process

It removes words that bear little or no content information such as articles, conjunctions, prepositions etc. Words which occur extremely often are also removed. Stemming It is a process of transforming word to its stem (normalize form). It builds basic form of words to identify word by its root. E.g. go is stem of gone, goes, going. Porter's Stemmer is the most popular algorithm and researchers make Text Mining: Process and Techniques Lata Gohil Text Mining: Process and Techniques 71 changes in the basic algorithm to cater to their requirements

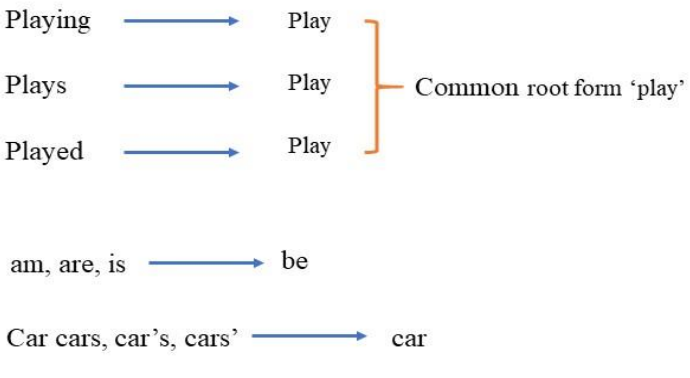


Fig. 3 Stemming process of text in text mining

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

It removes words that bear little or no content information such as articles, conjunctions, prepositions etc. Words which occur extremely often are also removed. Stemming It is a process of transforming word to its stem (normalize form). It builds basic form of words to identify word by its root. E.g. go is stem of gone, goes, going. Porter's Stemmer is the most popular algorithm and researchers make text mining. As there are many popular stemming algorithms like krovertz, porter , lovin's etc.

C. Sentence Similarity calculation

The key to summarization is conceptual similarity, not textual similarity. Similarity between features is common, for example, in natural language processing: words, n-grams, or syntactic n-grams can be somewhat different (which makes them different features) but still have much in common: for example, words "play" and "game" are different but related. When there is no similarity between features then our soft similarity measure is equal to the standard similarity. For this, we generalize the well-known cosine similarity measure with the following cosine similarity formula.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine Similarity is only calculated over non-NULL dimensions. When calling the function, we should provide lists that contain the overlapping items.

$$\left(similarity(A,B) = \frac{3 \cdot 10 + 8 \cdot 8 + 7 \cdot 6 + 5 \cdot 6 + 2 \cdot 4 + 9 \cdot 5}{\sqrt{3^2 + 8^2 + 7^2 + 5^2 + 2^2 + 9^2} \times \sqrt{10^2 + 8^2 + 6^2 + 6^2 + 4^2 + 5^2}} = \frac{219}{15.2315 \times 16.6433} = 0.8639 \right)$$

As per the above stated calculations, the different weights as per wordnet will be stored in a matrix which will be further used to calculate the similarity index. If its more than 50%, than the words of these sentences will be separated for further process where the similarity between words based on their meanings like synonyms and hyponyms is calculated.

D. WordNet

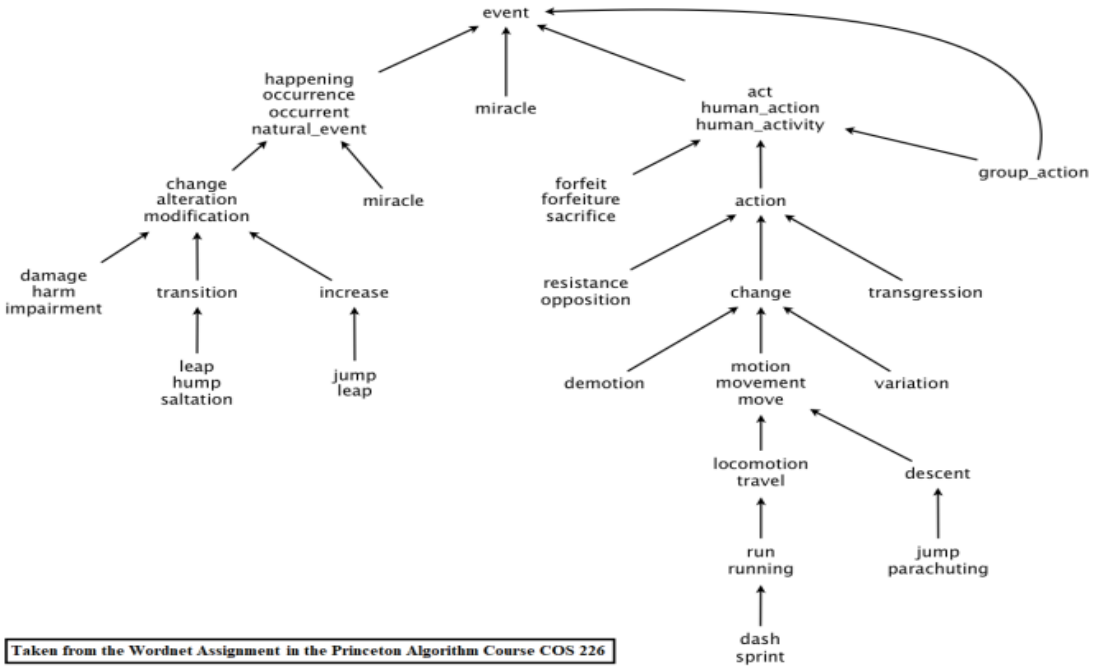
It's a lexical database for many popular languages. Corpora is being translated into many languages. It also aligns the words for their names, verbs and nouns properties. For example :

Proper names : Lady gaga, New-york city, House of John etc.

Dates : 1988, 2006, 2019 etc.

Application of WordNet:

- It supplies semantic meaning data for all valid dictionary words.
- Automatically detect words and don't fix the contexts.
- Can build corrective tools for editing.
- Finding neighboring words that are connected in WordNet.



Taken from the Wordnet Assignment in the Princeton Algorithm Course COS 226

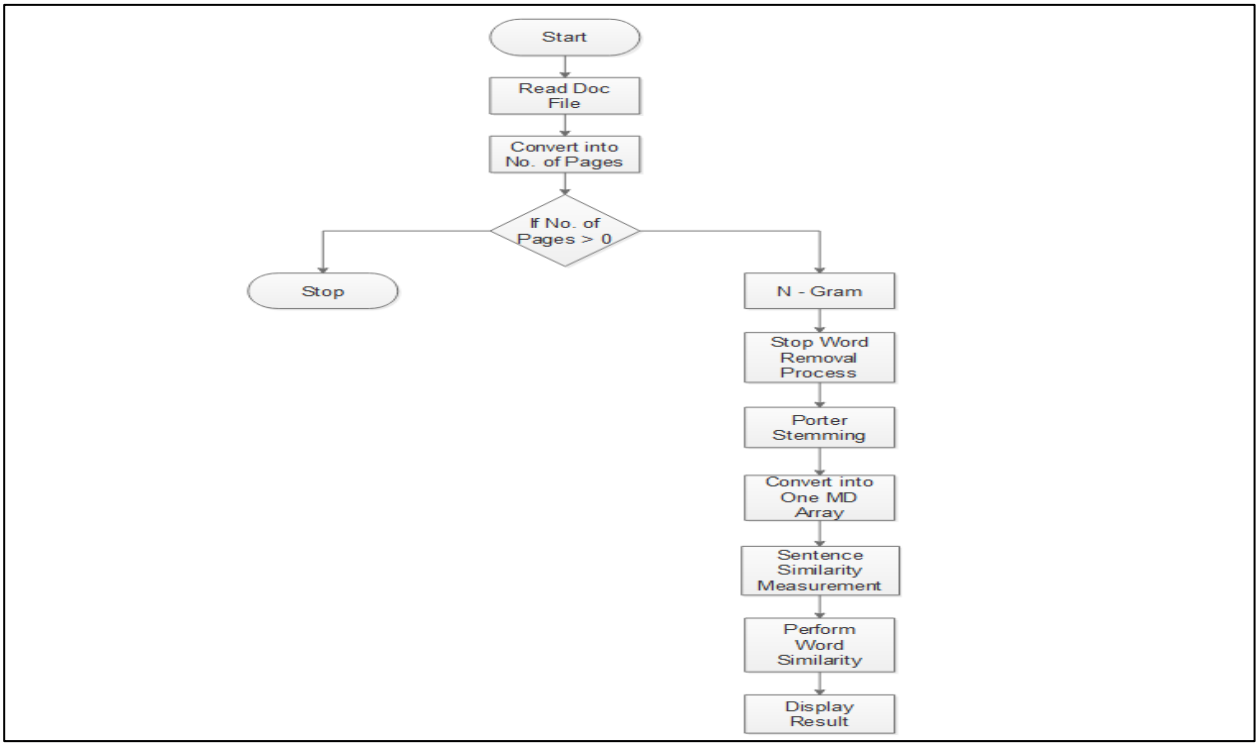
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

E. Word Similarity Check

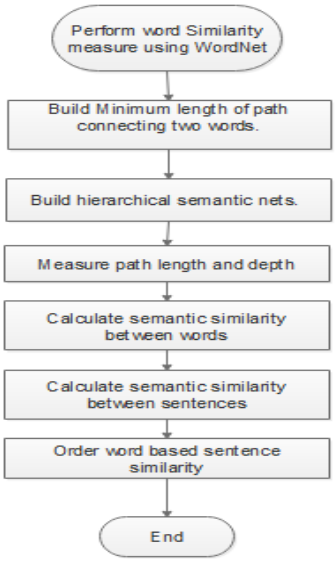
Word similarity check is a process to find the distance between two words or phrases. Based on the similarity index of sentences from a matrix, the words of theses sentences are first internally check and than the list of words which are not found similar will be added to a separate matrix. Semantic analysis processI helps users cluster similar articles by understanding the relatedness between different content and streamlines research by eliminating redundant text contents. Semantic analysis process has API that can help bloggers, publishing and media houses to write more engaging stories by retrieving similar articles from the past quickly, and news aggregators to combine similar news from different sources to reduce clutter in the feeds of their readers.

IV. PROPOSED WORK

In this section, proposed work flow is discussed in detail.



The above given flow diagram is for understanding the streams of NLP algorithms along with other text mining techniques. We have added the conditional approaches to decide the results and its effectiveness. The below given diagram is showing the sentence similarity and word similarity techniques we have used for novel approach. Here we have depicted the steps for adding words into a separate matrix which are having nearby or equal weigh thus making it confirm that no much similar or less related words are included.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

- Main aim of any research is to generate end results which are efficient than the earlier one. Manual assignment of quality keywords is error-prone, time-consuming and expensive. While extracting keywords, the methods and algorithms may perform differently.
- Automatic keyword extraction enables us to identify a small set of words, key phrases, keywords, or key segments from a textual document with the help of which the meaning of the document can be described. Outline of documentation is a productive and ground-breaking strategy that gives the short summary after effect of the entire information.

Proposed Algorithm

Input: Document.
Output: Keywords and summary

1. Identify number of pages.
2. Perform N-gram.
3. Perform Stop word Removal Process
4. Call Porter (Stemming).
5. Combine all different page words into one multidimensional array. (With picked sentences in a doc).
6. Perform word Similarity measure using WordNet
7. Display Result

V. RESULTS AND DISCUSSION

	F-measure		***Precision***		***Recall***	
	Base approach	Proposed Approach	Base Approach	Proposed Approach	Base approach	**Proposed Approach**
Noun	72.83	79.88	81.31	85.36	68.16	**73.21**
Verb and infinitive	**95.67**	**97.00**	**95.48**	**97.00**	**95.93**	**96.00**

Whole corpus measures the comparison based on precision, recall and f-measure on nouns and verbs from the whole document set. This difference in precision and recall of analyzed document counts is the result of average range of keywords having lower frequency across documents. The metric deg(w)/freq(w) favors average size keywords and therefore results in extracted keywords that occur in fewer documents in the whole Corpus is varied. As we can notice the difference between the results of previous research and the proposed research is approx 10 to 12% increasing in proposed approach. The noticeable difference found is in precision as its increasing more than 20% in proposed approach. These results will get accuracy by the following formula. Where the accuracy for noun is 78.81952 and for verb and infinitive is 96.49741.

Base paper Approach		
Sub-Corpus	Nouns	**Verbs and infinitives**
Fiction and compositions of pupils	76.15	**98.66**
Newspaper articles	77.35	**98.72**
Business and financial news	76.23	**98.57**
Computer texts	73.28	**96.32**
Legal texts	**67.82**	**88.05**

The results explore the results of different sub-corpus of base paper approach. The iterations for each word go on until we do not find next corpus or identifying new word is no longer cost-effective. The union criteria is therefore set to that the percentage of new words in all newly acquired key words in the mth iteration, rm, is very small, e.g., 1%. rm ¼ 1% indicates that we will need keyword based on the pre-defined incident dictionary (roughly 30 min to 1 hour in our experiments) in order to find one word, and the process is no longer cost-effective and should terminate.

Doc-set	**Precision**	**Recall**	**Accuracy**
Document set-1	85.48%	84.36%	85.00%
Document set-2	68.30%	70.00%	69.16%
Document set-3	85.00%	84.60%	85.30%
Document set-4	73.03%	75.05%	74.13%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

When keywords were discipline-specific, adjacency operators improved precision with little degradation in recall by proximity of terms may increase search success. The highest accuracy we have achieved is 85.30% in document set 3 , which consists of information about the text mining corpus. Here the least accuracy we have got is in document set 4 , which is a document consisting more equations and diagrams

VI. CONCLUSION

Keyword extraction is a powerful tool which enables us to scan large document collections efficiently. Automatic keyword extraction enables us to identify a small set of words, key phrases, keywords, or key segments from a textual document with the help of which the meaning of the document can be described. Text summarization is a useful technique for end user to supplement just required information in predetermined time. This paper contains the literature review about different techniques used to bring out keywords is largely depend upon methods on previously defined techniques for keyword generation; therefore text summarization method is greatly achieved based upon the keyword based techniques. We have used WordNet dictionary to find semantics of the words and phrases. apart from that we have added RDF and OWL to process the documents to build hierarchical semantic nets. We have received 10 to 15% better results in comparison with the previous approach where accuracy is improved by 10% and precision is improved with 15% in results. In future, we can consider using other formats of text and images to create an efficient summary approach.

REFERENCES

[1] Zoltán Subecz "Event detection in Hungarian texts with dependency and constituency parsing and WordNet" 2017 IEEE 14th International Scientific Conference on Informatics 10.1109/INFORMATICS.2017.8327276
[2] Akira Sasaki ; Junta Mizuno ; Naoaki Okazaki ; Kentaro Inui "Stance Classification by Recognizing Related Events about Targets" 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) , 10.1109/WI.2016.0100.
[3] De Wang ; Feiping Nie ; Heng Huang "Feature Selection via Global Redundancy Minimization" 2015 IEEE Transactions on Knowledge and Data Engineering , 10.1109/TKDE.2015.2426703.H.-L. Xu, X. Wu, X.-D. Li, and B. P. Yan, ``Comparison study of Internet recommendation system," J. Softw., vol. 20, no. 2, pp. 350_362, 2009.
[4] R. Chen, Q. Hua, Y.-S. Chang, B. Wei, L. Zhang, and X. Kong, ``A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks," IEEE Access, vol. 6, pp. 64301_64320, 2018.
[5] Kim, BD. & Kim, "A new recommender system to combine content-based and collaborative filtering systems" SO. J Database Mark Cust Strategy Manag (2001) 8: 244.
[6] Asif Salekin ; Hongning Wang ; Kristine Williams ; John Stankovic "DAVE: Detecting Agitated Vocal Events" 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 10.1109/CHASE.2017.74.
[7] Minh-Son Dao ; Noburu Babaguchi "Mining temporal information and web-casting text for automatic sports event detection" 2008 IEEE 10th Workshop on Multimedia Signal Processing, 10.1109/MMSP.2008.4665150.
[8] Lei-Lei Shi ; Lu Liu ; Yan Wu ; Liang Jiang ; James Hardy "Event Detection and User Interest Discovering in Social Media Data Streams" 2017 IEEE Access, 10.1109/ACCESS.2017.2675839.
[9] J.N.Madhuri ; Ganesh Kumar.R, " Extractive Text Summarization Using Sentence Ranking" IEEE 2019 International Conference on Data Science and Communication (IconDSC)10.1109/IconDSC.2019.8817040
[10] ShivangiModi ; RachanaOza , " Review on Abstractive Text Summarization Techniques (ATST) for single and multi documents" IEEE 2018 International Conference on Computing, Power and Communication Technologies (GUCON)10.1109/GUCON.2018.8674894
[11] P. Krishnaveni ; S. R. Balasundaram , " Automatic text summarization by local scoring and ranking for improving coherence" IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC)10.1109/ICCMC.2017.8282539
[12] Santosh Kumar Bharti, KorraSathyaBabu , " Automatic Keyword Extraction for Text Summarization: A Survey" 2017arXiv:1704.03242
[13] Changjian Fang, Dejun Mu, Zhenghong Deng, Zhiang Wu "Word-Sentence Co-Ranking for Automatic Extractive Text Summarization" 2016 Elsevier Ltd. 10.1016/j.eswa.2016.12.021
[14] AlokRanjan Pal, DigantaSaha "An Approach to Automatic Text Summarization using WordNet" IEEE, 2014 10.1109/IAdCC.2014.6779492
[15] www.Researchgate.net
[16] Shohreh Rad Rahimi, MohamadAbdolahi, AliToofanzadehMozhdehi "An Overview on Extractive Text Summarization" IEEE, 2017 10.1109/KBEI.2017.8324874
[17] Deepak Sahooa, AshutoshBhoib, Rakesh Chandra Balabantarayc "Hybrid Approach To Abstractive Summarization" Elsevier Ltd., 2018 10.1016/j.procs.2018.05.038