Hybrid Approach to Generate Summary by Extracting Keywords in Education Domain with Text mining Techniques

[1]Apexa Bhavsar,[2]Kinjal Thakar
[1]M.E. Student,[2]Assistant Professor
[1]Department of Information Technology,
[1]Silver Oak College of Engineering & Technology, Ahmedabad, India

Abstract : As of recent studies, information is developing quickly in each area, for example, news, social media, banking, education and many more. Because of the excessive amount of information, there is a need of programmed algorithm which will be skilled to summarize the information particularly text based information from unique record without losing any basic purposes and meanings. "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 hybrid approach where keywords and summarization of document could be automatically generated using text mining techniques, since text summarization process is highly depend on keyword extraction.

IndexTerms **- Text Mining, Machine learning, Natural language processing, WordNet, Keyword Extraction, Text Summarization**

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.

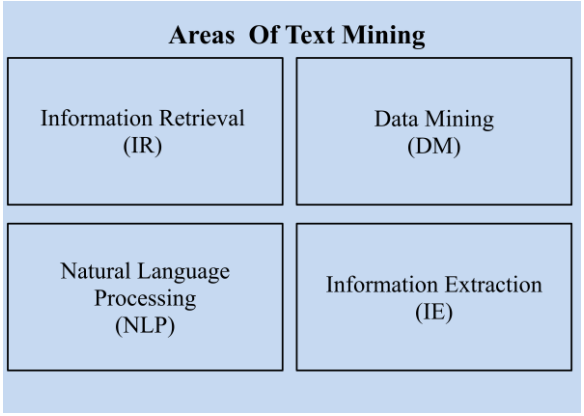

Fig. 1. Areas of text mining

2. Data Mining (DM)

Searching for examples in information.
Look databases for covered up and obscure examples.
3. Natural Language Processing (NLP)
It is the investigation of human dialect.
Convey the framework in the data extraction stage as an information.
PCs can comprehend common dialects as people do.
4. Information Extraction (IE)
Undertaking of naturally separating organized data from unstructured. incorporates preparing human understandable messages by methods of NLP.

Searching for examples in information.
Look databases for covered up and obscure examples.
3. Normal Language Processing (NLP)
It is the investigation of human dialect.
Convey the framework in the data extraction stage as an information.
PCs can comprehend common dialects as people do.
4. Information Extraction (IE)
Undertaking of naturally separating organized data from unstructured.
include preparing human understandable messages by methods for NLP.

TEXT MINING PROCESS
A process of Text mining involves a series of activities to be performed to mine the information [2]. These activities are:

1. Text Pre-processing
i. Text Cleanup
Text Cleanup means removing any unnecessary or unwanted information.

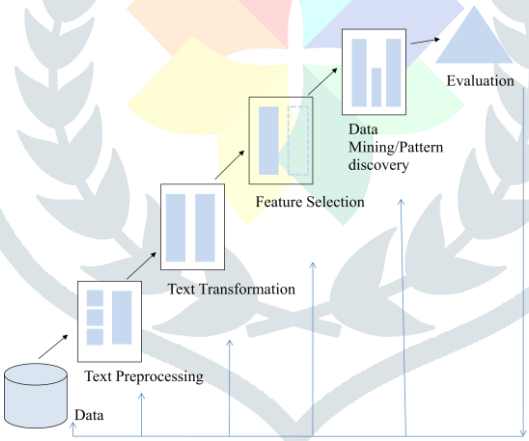

Fig. 2. Text mining process

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
Evaluate the result.
6. Applications
Text Mining can be applied in a variety of areas. Web mining, Medical, Resume Filtering etc.

WHY TO EXTRACT KEYWORDS AUTOMATICALLY?

Programmed keyword extraction is the way toward choosing words and expressions from the content record that can, best case scenario venture the center conclusion of the archive with no human intercession relying upon the model . The objective of programmed slogan extraction is the utilization of the power and speed of current algorithm capacities to the issue of access and recovery, worrying upon data association without the additional expenses of human annotators.
Web based metadata and text based values have different methods for keyword extraction process.

WHY TO EXTRACT TEXT SUMMARY AUTOMATICALLY?

Reducing text and size of documents in such a way that important text keep itself till end. The Summarization is where the most essential highlights of a content are separated and arranged into a short conceptual of the required textual information. Summary generated are generally around 17% of the first content but then contain everything that could have been gained from perusing the first record. It requires accurate crafting of text. It performs equally well or less well across various domains.

Text summarization is a productive and ground-breaking system to give a look at the entire information. The content outline can be accomplished in two different ways to be specific, abstractive rundown and extractive synopsis. The abstractive synopsis no standard algorithm is used. These rundowns are gotten from realizing what was communicated in the report and after that changing over it into a shape communicated by the PC. It looks like how a human would condense the report subsequent to understanding it. Though, extractive rundown remove subtle elements from the first archive itself and present it to the per user. [4]

II. RELEVANT WORK

A few papers are worried about tag based extraction and rundown.

In this paper [5], they have structure and break down a group strategy to naturally remove such words from single report. Methodologies utilized are: RAKE, TAKE, and TEXTRANK. Stemmer function is than added that can efficiently extract root of words. Tried the exactness and gotten generally speaking better execution utilizing hybrid approach. Process is-RAKE TAKE utilize fox-stop rundown and Text rank use NLTK library. Apply Filtering heuristic at that point and then recalculate scores is a novel methodology adapted in this paper. Apply Dynamic edge capacities. Constraint is with "precision". Extract watchwords from Single document. Get Results just with informational index of "Software engineering and data Technology".

In this paper [6], they have present fundamental improvement including new algorithm for watchword extraction and synopsis age at the same time over a subset of reports. A Novel technique for keyword extraction and different report outlines, Provide reflection of archives and fabricate connections among records.

Various docs can have distinctive kind of information which is considerable. C value technique, FGB algorithm were utilized. Apply algorithm just for "online wellbeing networks". Some positioning approach can be useful to yield better outcomes.

In this paper [7], they proposed the programmed watchword extraction framework and Thai site order framework which can consequently refresh the word reference and arrange site in Thai.

SVM algorithm, Naïve Bayes ,Decision tree, administered learning algorithm and semantic methodology with TF-IDF weighting system were utilized. Spotlight on algorithm speed with adequate precision. Not need of a specialist. Utilization of WordNet for thai words and sentences is proposed.

In this paper [8], they investigate distinctive keywords and key phrase extraction algorithms for the space of online protection approaches. To do this they have utilized a variety of surely understood strategies, for example, TF-IDF, RAKE, Text Rank, and Alchemy API, benchmarked against manual annotation.
In this paper the required stages of preprocessing of text from documents have been performed primarily. Than for the post processing stage, they have used the novel techniques to generate keywords. Limitation of this paper is-Results influenced by four noteworthy sorts of errors: Over age blunders, Redundancy Errors, Infrequency blunders and Evaluation Errors.

In this paper [9], they proposed a catchphrase based suggestion framework (KBRS), where the client's inclinations are demonstrated by watchwords for web based life. They utilized a client based community oriented separating (UCF) calculation to

give suggestions. The required result is to aim at better practiced and more flexible and adaptable execution of proposed algorithm with Hadoop framework, utilizing the MR framework. Distinguish User inclinations, Compute Similarity, Generate Recommendation list, Collaborative separating calculation.

In this paper [10], they proposed a machine learning based methodology with rich list of capabilities that consequently recognize occasions with consistency and reliance parsing and WordNet for Hungarian texts. Technology-NLP, score Corpus, Weka information mining suit, WordNet double grouping. Got Best outcome on News paper article and Worst outcome on lawful writings.

In this paper [11], they identified and Detect 5 critical verbal forms: requesting encourage, verbal advances, questions, reviling, chatting with tedious sentences. Innovation TF-IDF Features, Language Model Features, Smoothing strategies. Precisely and naturally distinguishes the 5 vocal occasions of the Cohen-Mansfield stock.

III. RESEARCH GAP

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 f-measure 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].

IV. PROPOSED WORK

In this section, method used for our proposed work is discussed in detail.

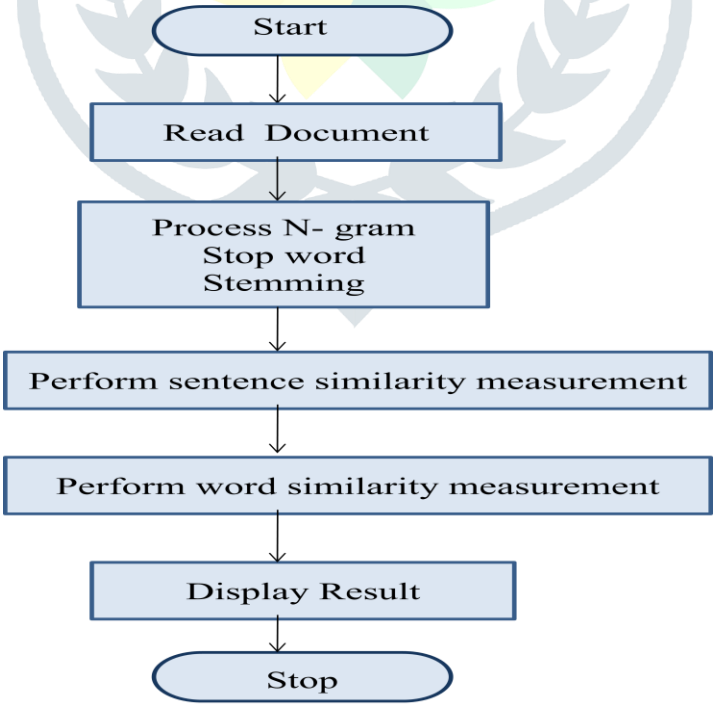

Figure 3. Proposed Flow

PROPOSED ALGORITHM

Algorithm Automatic keyword extraction and Summarization

Input- Take document (Education Domain)
Step 1 - Identify no. of pages.
Step 2- Perform N-gram (Hybrid with Karp-rabin String matching Algorithm).
Step 3 - Perform Stop word Removal Process.
Step4 - Call Porter (Stemming).
Step 5 - Combine all different page words into one multidimensional array. (with picked sentences in a doc).
Step 6 - Perform word Similarity measure using WordNet.

- Build Minimum length of path connecting two words.
- Build hierarchical semantic nets. (Using RDF and OWL).
- Measure path length and depth.
- Calculate semantic similarity between words.
- Calculate semantic similarity between sentences.
- Order word based sentence similarity.

Step 7 – Display Result.

The goal of our approach is to extract keywords automatically and summary for education domain. Therefore First, We need to identify no. of pages of the document. Need to perform **N-gram Algorithm-"Gram"** derived from Greek which means **"letter".** A set of co-occurring words within a given window When computing the n-grams typically move one word forward. Use to develop language model.

Develop features for supervised Machine Learning models. Perform **Stop word Removal process**- A set of commonly used words in any language. Commonly eliminated from many text processing applications because these words can be distracting, non-informative and are additional memory overhead**. Perform Stemming - Call Porter().**
Step 1 - Remove plurals.
Step 2 - If there is another vowel in the stem,
Step 3 - Double Suffixes in single ones –examining second to last letter.
Step 4 - Examining last letters.
Step 5 - Examining second to last letters continued.
Step 6 - Tidying up.
Step 7 - Return.

Perform word similarity measure using WordNet, Build minimum length and path, measure path length and depth, calculate semantic similarity between words and sentences. Get Result.

V. RESULTS

TABLE I. RESULTS ON THE WHOLE CORPUS (%)

	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.

$$f = 2*(precision * recall) / (precision + recall)$$

TABLE II. RESULTS ON THE SUB-CORPORA
(F-MEASURE, %)

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 m^{th} 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.

TABLE II

Proposed Approach			
	Precision	***Recall***	***Accuracy***
Document set-1	85.65%	84.36%	85.0001
Document set-2	68.35%	70.00%	69.1651
Document set-3	85.00%	86.00%	85.4970
Document set-4	70.35%	75.65%	72.9038
Document set-5	73.25%	75.05%	74.1390

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 85.48% 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 are hopeful to get good results in comparison with the previous one.

REFERENCES

[1] https://data-flair.training/blogs/text-mining/
[2] http://shodhganga.inflibnet.ac.in/bitstream/10603/34713/9/09_chapter%201.pdf
[3] https://www.springer.com/gp/authors-editors/authorandreviewertutorials/writing-a-journal-manuscript/title-abstract-and-keywords/10285522
[4] Data Mining_ The Textbook [Aggarwal 2015-04-14]
[5] Tayfun Pay, Stephen Lucci "Automatic Keyword Extraction: An Ensemble Method" in big data, 2017 IEEE International Conference on Big Data (BIGDATA),IEEE 2017, pp. 4816-4818.
[6] Weijia Xu ,Wei Luo, Nicholas Woodward, Yan Zhang "Supporting Data Driven Access through Automatic Keyword Extraction and Summarization", 2015 IEEE International Congress on Big Data ,IEEE 2015, pp 704-707.
[7] Adsadawut Chanakitkarnchok, Kulit Na Nakorn, Kultida Rojviboonchai "Automatic Keyword Extraction System for Thai Website Categorization System", 2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) ,IEEE 2017, pp 206-209.
[8] Dhiren A. Audich, Rozita Dara, Blair Nonnecke, "Extracting Keyword and Keyphrase From Online Privacy Policies", The Eleventh International Conference on digital Information management (ICDIM 2016), ICDIM 2016, pp 127-132.
[9] Sandra Elizabeth Salim, R. Jebakumar, "KBRS - Keyword Based Recommendation System in Social Networks ",2015 International Conference on Innovation Information in Computing Technologies(ICIICT),Chennai, India, ICIICT 2015.
[10] Zoltán Subecz, "Event Detection in Hungarian Texts with Dependency and Constituency Parsing and WordNet", 2017 IEEE 14th International Scientific Conference on Informatics, IEEE 2017, pp 365-371
[11] Asif Salekin1, Hongning Wang1, Kristine Williams2, John Stankovic1salekin, "DAVE: Detecting Agitated Vocal Events",2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), IEEE 2017, pp 157- 166.
[12] https://catalog.data.gov/dataset?groups=education2168
[13] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents." Text Mining., pp. 1–20, 2010.
[14] T. Pay, "Totally automated keyword extraction," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE,2016, pp. 3859–3863.
[15] C. Zhang, "Automatic keyword extraction from documents using conditional random fields," Journal of Computational Information Systems, vol. 4 (3), 2008, pp. 1169-1180.