

**Thesis No: CSER-M-18-06**

**A STUDY ON KNOWLEDGE EXTRACTION FROM OFFICIAL BANGLA  
DOCUMENTS**

By

**Monika Gope**



Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

December, 2018

# **A Study on Knowledge Extraction from Official Bangla Documents**

By

**Monika Gope**

Roll No: 1207554

A thesis submitted in partial fulfillment of the requirements for the degree of  
Master of Science in Computer Science and Engineering



Department of Computer Science and Engineering

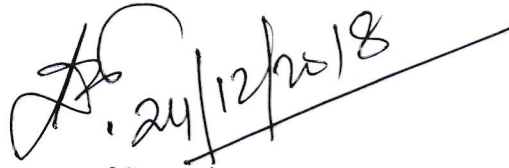
Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

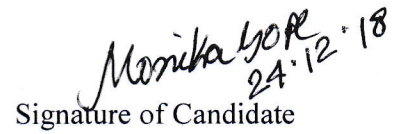
December, 2018

## Declaration

This is to certify that the thesis work entitled “**A Study on Knowledge Extraction from Official Bangla Documents**” has been carried out by Monika Gope in the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

A handwritten signature in black ink, appearing to be 'A. 24/12/2018', written over a diagonal line.

Signature of Supervisor

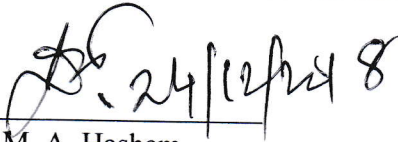
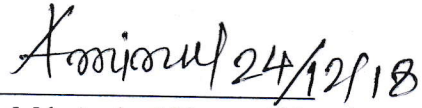
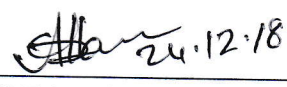
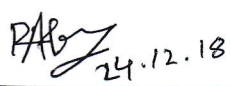
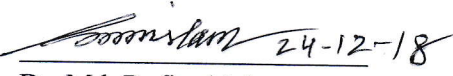
A handwritten signature in black ink, appearing to be 'Monika Gope 24.12.18', written over a diagonal line.

Signature of Candidate

## Approval

This is to certify that the thesis work submitted by Monika Gope entitled “A Study on Knowledge Extraction from Official Bangla Documents” has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering in the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in December, 2018.

### BOARD OF EXAMINERS

1.   
 Dr. M. M. A. Hashem  
 Professor, Department of Computer Science and Engineering  
 Khulna University of Engineering & Technology, Khulna-9203. Chairman  
(Supervisor)
2.   
 Dr. Md. Aminul Haque Akhand  
 Head of the Department  
 Department of Computer Science and Engineering  
 Khulna University of Engineering & Technology, Khulna-9203. Member
3.   
 Dr. K. M. Azharul Hasan  
 Professor, Department of Computer Science and Engineering  
 Khulna University of Engineering & Technology, Khulna-9203. Member
4.   
 Dr. Kazi Md. Rokibul Alam  
 Professor, Department of Computer Science and Engineering  
 Khulna University of Engineering & Technology, Khulna-9203. Member
5.   
 Dr. Md. Rafiqul Islam  
 Professor, Computer Science and Engineering Discipline  
 Khulna University, Khulna. Member  
(External)

## **Acknowledgment**

All the praise to the almighty Lord, whose blessing and mercy succeeded me to complete this thesis work fairly. I gratefully acknowledge the valuable suggestions, advice and sincere co-operation of Dr. M. M. A. Hashem, Professor, Department of Computer Science and Engineering, Khulna University of Engineering & Technology, under whose supervision this work was carried out. His open-minded way of thinking, encouragement and trust makes me feel confident to go through different research ideas. From him, I have learned that scientific endeavor means much more than conceiving nice algorithm and to have a much broader view at problems from different perspectives. I would like to convey my hearty ovation to all the faculty members, officials and staffs of the Department of Computer Science and Engineering and IICT as they have always extended their co-operation to complete this work. I am extremely indebted to the members of my examination committee for their constructive comments on this manuscript. I would also like to thank my parents for their wise counsel. Last but not least, I wish to thank my friends and registrar office of KUET for their constant support.

**Author**

## **Abstract**

Bangla is the seventh largest spoken language in the world. However, the information searching in the digital Bangla papers is a tiresome job as its ends up with incorrect and a very few information. It is difficult because wide computational resources for Bangla are very limited. It is literally infeasible to list and analyze the Bangla data manually. Several approaches for identifying and extracting tables, figures, emotion, reviews, and algorithms have been done in English. Furthermore, knowledge extraction in Bangla documents for emotion or opinion detection and sentence extraction for summarization have been explored. However, they do not provide enough textual information for the user for Bangla text content. In this work, we proposed a domain specific composite approaches to find out the agendas, and its decisions from the minutes of meeting of academic council of Khulna University of Engineering & Technology (KUET) with Query-based features, Content-based features, Context-based features and Semantic features. We also demonstrated the techniques to categorize the knowledge where a single query is given by the user and displayed the result sequentially by date. All the operations are presented with sufficient theoretical analysis and experimental results.

## Contents

	<b>PAGE</b>
Title Page	i
Declaration	ii
Approval	iii
Acknowledgment	iv
Abstract	v
Contents	vi
List of Tables	viii
List of Figures	ix
<b>CHAPTER I Introduction</b>	<b>1</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Problem Statement	3
1.4 Objectives	3
1.5 Scope	4
1.6 Contributions of the Thesis	4
1.7 Organization of the Thesis	5
<b>CHAPTER II Literature Review</b>	<b>6</b>
2.1 Introduction	6
2.2 The Realization of Extracting Elements from Documents	6
2.2.1 OCR-based Analysis of Mathematical Texts from PDF	6
2.2.2 Extraction of PDF Information	7
2.2.3 Detection and Segmentation of Table of Contents	9
2.2.4 Extracting Metadata Information	7
2.2.5 Extracting Bibliography	11
2.2.6 Extraction of Data Points and Text Blocks	12
2.2.7 Summarizing Figures, Tables, and Algorithms in Scientific Publications	13
2.2.8 Extracting Algorithms in Scholarly Big Data	15
2.3 The Realization of Extracting Bangla Text, Image, Number and Knowledge	15
2.3.1 Bangla Number Extraction and Recognition from Document Image	16
2.3.2 Phrase-level Polarity Identification for Bangla	16
2.3.3 Bangla Text Extraction from Natural Scene Images	17
2.3.4 Sentiment Analysis on Bangla and Romanized Bangla Text	17
2.3.5 Bangla Text Summarization by Sentence Extraction	18
2.4 The Realization of Extracting Keywords	18
2.4.1 Rapid Automatic Keyword Extraction	19
2.4.2 Other Schemes	20

2.5 Discussion	20
<b>CHAPTER III Theoretical Consideration</b>	<b>21</b>
3.1 Introduction	21
3.2 Word Density	21
3.3 Similarity with Caption	22
3.4 Naive Bayes Classifier	23
3.5 Sentence Selection	24
3.6 Confusion Matrix	24
3.7 Graph-Based Centrality and PageRank and TextRank	26
3.8 Mixture Models and EM Algorithm	27
3.9 Discussion	28
<b>CHAPTER V Methodology</b>	<b>29</b>
4.1 Introduction	29
4.2 Realization of the Method for Proposed Bangla Knowledge Extraction	29
4.3 Data Selection and Pre-processing	32
4.3.1 Selection of the Target Data	33
4.3.2 Pre-processing the Data	33
4.4 Feature and Patterns Specifications for Decision Extraction	34
4.4.1 Query-Based Features	34
4.4.2 Content-Based Features	36
4.4.3 Context-Based Features	37
4.5 Ordering the Documents Chronologically	40
4.6 Processing of the Keywords from the Extracted Decision Pool	40
4.7 Feature and Pattern Extraction for Decisions with User Query	42
4.7.1 Content-Based Features	42
4.7.2 Semantics Features	43
4.7.3 Context-Based Features	44
4.7.4 Classify the Documents with Keywords	44
4.8 Conclusion	45
<b>CHAPTER V Results and Discussions</b>	<b>47</b>
5.1 Experimental Setup	47
5.2 Performance Analysis of the Structure	47
5.2.1 Extraction of Agenda And Decisions Text Analysis	47
5.2.2 Finding User Query from the Extracted Decision Pool Analysis	51
5.3 Discussion	53
<b>CHAPTER VI Conclusions</b>	<b>55</b>
5.1 Summary	55
5.2 Recommendations for Future Works	56
<b>References</b>	<b>57</b>



## LIST OF TABLES

<b>Table No.</b>	<b>Description</b>	<b>Page</b>
2.1	Features Used in Book Metadata Extraction	11
2.2	Rules for Generating Venue Alias	12
2.3	A Grammar for Document-Element Captions	14
5.1	Precision, Recall and F1 for “Decision” Detection for Random 8 Documents	48
5.2	Total Set of Data with Precision, Recall and F1	49
5.3	Total Set of Data for One Keyword – মেকানিক্যাল with Precision, Recall and F1	53

## LIST OF FIGURES

Figure No.	Description	Page
1.1	Knowledge Extraction Process	2
2.1	Decision Tree to Identify Two Types of Content Pages: TOC-I and TOC-II.	10
2.2	Decision Tree for TOC Segmentation	10
3.1	Data, Information, Knowledge, Wisdom Chain	21
3.2	Confusion Matrix	24
3.3	Confusion Matrix for Total Predicted Positive	25
3.4	Confusion Matrix for Actual Positive	26
4.1	Design of Proposed Knowledge Extraction from Official Bangla Documents	30
4.2	Proposed Knowledge Extraction Algorithm	30
4.3	Design of Agenda and Decision Extraction from Official Bangla Documents	31
4.4	Proposed Algorithm for Agenda and Decision Extraction from Bangla Documents	31
4.5	Design of Keyword Extraction and Finding User Query from Documents	32
4.6	Proposed Algorithm for User Query Extraction with Features from Decision Pool	32
4.7	Example 4.1	35
4.8	Decision Making Phrases	37
4.9	Example 4.2	38
4.10	Example 4.3	39
4.11	Example 4.4	39
4.12	Example 4.5	40
4.13	Stopwords for the Domain	41
4.14	Example of the Frequency of Two Words :Post Facto And Mechanical	42
4.15	Connection Word List	45
4.16	Example of Occurrence of Words	45
5.1	Precisions of the Three Methods of Random 8 Documents	49
5.2	Total Decisions Detected of the Methods- A) Content-Based Method and B) Context-Based Method Merge with Content-Based Method and C) Total Decision Counted Manually for 29 Documents	50
5.3	BM25 Score and Sentence Weight (Query Word- “মেকানিক্যাল”)	50
5.4	Example of “Decision” Detection in a Single Document	51
5.5	Top-Ranked Sentences by Textrank	52

<b>Figure No.</b>	<b>Description</b>	<b>Page</b>
5.6	Gaussian Curve for Three Clusters	52
5.7	K-Means of the Two Clusters	53

## CHAPTER I

### Introduction

#### 1.1 Introduction

The volume of data being created and warehoused is rising exponentially, due in great part to the ongoing progresses in computer technology [1]. It is estimated by the experts that approximately 2.5 trillion PDF generated each year all over the world, contributing to every segment of the global economy [2]. Nevertheless, content coded in PDF is condensed to streams of printing commands to present a visual draft with text, images, tables, graphs etc. [3]. As a result a significant number of Bangla digital documents, such as reports, agenda, paper, journals, dealings developed by various officials of different Govt. and non-Govt. organization, is on the rise. To unlock the information embedded within this data that surround us, introduces new challenges [1]. To solve this problems various data mining techniques are developed and also faces various challenges.

Data mining is a procedure that takes data as input and outputs knowledge [1]. Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [4]. For data mining process, the first few steps involve preparing the data where the relevant data must be selected from a potentially large and diverse set of data, any necessary preprocessing must then be performed, and finally the data must be transformed into a representation suitable for the data mining algorithm that is applied in the data mining step [1].

The surge of newer and newer dealings, resolution, and agendas with important decisions in Bangla PDF makes it infeasible to list and analyze the data manually. For any kind of decision making in office, research, business, the previous data are very significant as it provides the rules, policies and various decision, solution, problems etc. To find any particular information from this huge data store such as Big Data is literally very difficult. Substantially the information searching in these digital papers is a nontrivial job. The information searching in the digital Bangla papers is a tiresome job as its ends up with

incorrect and a very few information. It is difficult because wide computational resources for Bangla are very limited. Bangla document analysis and knowledge extraction are literally very difficult.

Precisely, in this research work, we are going to propose a technique which will discover and extracts agenda and decisions which are taken in official meetings of Khulna University of Engineering & Technology (KUET) and find out the exact knowledge searched by the user from digital Bangla official minutes of meeting of KUET.

## 1.2 Motivation

While working with Bangla documents, we didn't find any option to analyze the text and there was no way to search the information from a set of Bangla PDF files. To gather knowledge manually from these kind of pools, become so boring and tiring. We were looking for some decision making statements on particular topics from the minutes of meeting of academic council of Khulna University of Engineering & Technology (KUET). But it was so difficult to find the exact knowledge as there was no proper system to solve the problem. For these particular problem domain, there is a need for a particular problem representation technique, and/or a particular solution technique with the Bangla minutes of meeting of KUET.

Therefore we want to make an effort to solve the problem for assisting the user to extract knowledge for a specific domain of Bangla resolutions of academic council of KUET according to his needs. Motivated by [1] and [4] we have used the following process to solve our problem which is shown in Fig 1.1.

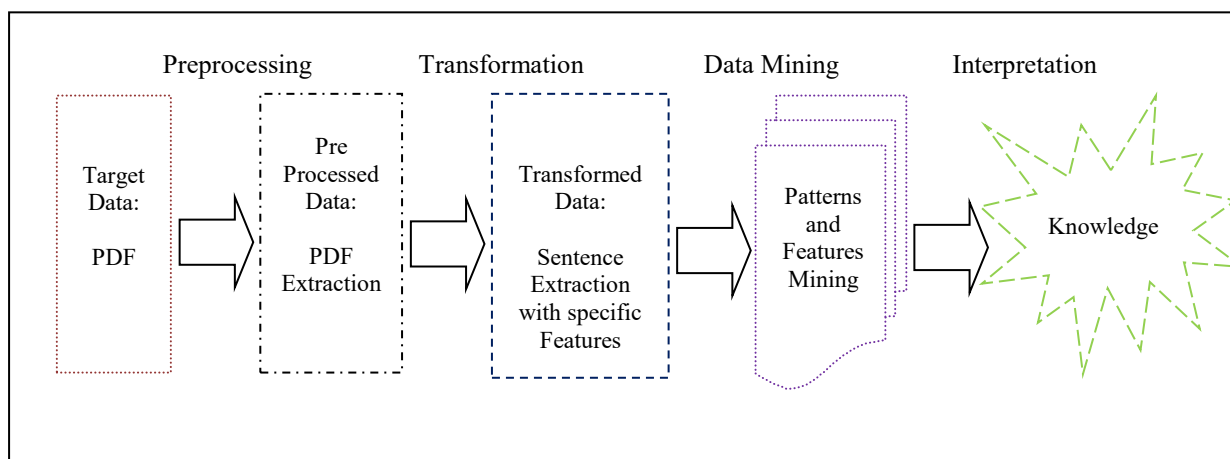


Fig 1.1: Knowledge Extraction Process

### 1.3 Problem Statement

Identifying and extracting informative entities such as mathematical expressions [5], [6], tables of contents [7], [8], figures [9], [10], from documents have been studied widely. Bhatia, et al., recommended a set of methods used for detecting document-elements with captions, e.g. tables, figures, and pseudocodes [11], [12]. Knowledge extraction in Bangla documents for example emotion or opinion detection from Bangla Blogs and News, Bangla text extraction from natural images, Bangla sentiment analysis from micro-blogs, news portal and product review portal, number extraction are extensively explored [13], [14], [15]. A work on Bangla sentence extraction for text summarization [16] is also studied earlier.

However, none of these procedures can really extract the information poised in the Bangla official document and therefore do not serve our purpose the specific domain. Furthermore, they do not provide enough textual information for the user.

We want to propose a domain specific composite approaches to find out the agendas and its decisions chronologically from the resolutions or minutes of meeting of an official Bangla documents based on various specifications as Content-based features (Query-based or Phrase mapping, Similarity calculation with the Phrase and Sentence scoring) and Context-based features (surroundings factors of the sentence and the location of the sentence) and also want to find out the knowledge where a single query is given by the user and display the result sequentially by date with Content and Semantic based method.

### 1.4 Objectives

The traditional approaches are unable to extract the Bangla knowledge such as decisions and user query from the PDF files and categorize it accordingly proficiently. To cope with this situation, the scientists have appreciated keywords extraction and rules extractions for English or other languages.

Therefore, main objectives of this research topic can be summarized as –

- To find out the proper needs of domain specific Bangla text content in PDF
- Query, Content, Context and Semantic based feature Specifications and Extraction for Agendas, decisions and user query of official Bangla documents

- Create a knowledge base for official Bangla documents and categorize the dataset with the keywords.
- To find out exact and quick knowledge discovery from a surge of Bangla PDFs from a user query.

### 1.5 Scope

The proposed model with the data set is experimented under these scope:

- We have selected the resolutions of the academic council of Khulna University of Engineering & Technology (KUET) as a domain for a case study where data are stored as Bangla PDF.
- We have experimented with 29 resolutions of Academic Council's meeting of KUET for Decision extraction and user query extraction.
- Here the PDF should be made from a Unicode file and the font should be embedded with the file.
- Computation can be done independently in each converted PDF and the data set is very small for big data analysis
- The rule base and knowledge base is for a specific domain with a small dataset and do not extract information from the tabular data.

All experiments in done on a windows machine and we used Java and Python as programing languages to implement our algorithm and Weka [17] as an implementation tool. We have used nltk and other python packages. Naïve-Bayes Classifier model and Gaussian is used to classify the result.

### 1.6 Contributions of the Thesis

This paper has the following vital contributions:

- We have proposed domain specific composite approaches to find out the agendas and its decisions chronologically from the resolutions of the Academic Council of KUET based on various specifications as Content-based features (Phrase mapping, Similarity calculation with the Phrase or Query and Sentence score) and Context-based features (surrounding factors of the sentence and the location of the sentence) and the result is presented sequentially by date with Content-based method.

- We have proposed composite approaches to find out the decisions from the decision pool chronologically from the resolutions of the Academic Council of KUET based on Query-based, Content-based, Semantics-based, Context-based features of the sentences.
- We have also revealed the techniques to categorized and classify the decisions from the documents using keywords extraction.

### **1.7 Organization of the Thesis**

- **Chapter II** presents Literature Review of the similar domains and finds some limitations of the existing works.
- **Chapter III** presents theoretical considerations of the background of the work.
- **Chapter IV** proposes the Model. The chapter also describes different rules and algorithm with examples.
- **Chapter V** shows the experimental results of the proposed scheme with discussion.
- **Chapter VI** exhibits the future direction of the proposed model and outlines the conclusions.



## CHAPTER II

### Literature Review

#### 2.1 Introduction

The intensive expansion of the internet and electronic publishing has led to an enormous amount of scientific documents being accessible to users, but, they are typically unreachable to those with visual deficiencies and often partly compatible with software and hardware such as tablets and e-readers [5]. However, several approaches have been proposed for extracting elements like figures, tables and useful information from the digital documents.

#### 2.2 The Realization of Extracting Elements from Documents

The knowledge extraction paradigm is prevalent in most sciences and it has drawn consideration from the data mining research community for several years. Some of the related research are given below:

##### 2.2.1 OCR-based Analysis of Mathematical Texts from PDF [5]

Document analysis of mathematical texts is a confounding problem for digital documents in regular formats in the context of PDF documents. Some uses an OCR approach for character identification together with a virtual link network for structural investigation. The other uses straight extraction of symbol information from the PDF file with a two stage parser to extract layout and expression constructions. Through reference to ground truth data, [5] contrast the efficacy and correctness of the two methods with respect to character identification and structural analysis of mathematical expressions with respect to layout analysis.

OCR-based scheme of PDF [5] is applied in the Infty system [18] and uses that system's services for identifying mathematical texts from scanned documents. That is, it primary evinces the PDF document into an image before finishing layout analysis, segmentation

and character and mathematical expression recognition. Great recognition rates can be obtained with digital PDF documents as they are moderately free of noise, consequently less disposed to usual recognition mistakes. Structural elements that Infty can detect are; titles, headings, author information, headers, footnotes, page numbers and mathematical components. Headers and footers are recognized by having smaller than typical sizes and appearing at the top or bottom of the page.

### **2.2.2 Extraction of PDF Information [5]**

OCR method has the benefit that the engine is applied to a noise free image created from the PDF version, but it fails to apply any of the information obtainable from the PDF document. Extraction method [5] targets to do accurately this, by extracting information on characters, their fonts and sizes with their precise location in the document. This information is then exploited to collect the main document with a specific importance on the mathematical expressions. These procedures are implemented in the dedicated PDF extraction tool Maxtract [5], which implements a linear grammar tactic for identifying mathematical expressions [19] with font and size information on characters for augmented recognition [20]. For this assessment the tool was considerably prolonged to not only work on by hand clipped mathematical expressions only but automatically on complete PDF documents with layout analysis and segmentation of mathematics and text. Primarily, all characters on a given page with their particular placing have to be extracted. Unluckily, PDF documents do not encompass the true bounding box information about the characters that they contain. As an alternative, they identify the point where the characters are rendered to on the page and deliver only a very simple bounding box guesstimate for individual character.

To obtain the precise bounds necessary, the PDF document is rendered to a 600dpi TIFF bitmap image, with the bounding boxes of each glyph (i.e., connected component) in the page extracted then registered with the character information from the original PDF. For most characters this is a simple matter of translating and scaling the coordinates obtained from the PDF and matching them to the corresponding glyphs.

The extraction yields an exact list of characters on the PDF page, their bounding boxes and font and size information. Layout analysis was necessary in order to recognize lines and columns. An analyzed PDF page contains number of lines, each with an overall

bounding box and list of symbols, with character and glyph information. Each extracted line is then analyzed separately to rebuild its spatial layout. In a first parsing step characters are clustered exploiting both the extracted information on their size and font as well as spacing information calculated from their bounding boxes. Characters are clustered together if they are either alpha characters or single digits and they

- a) Use the identical font,
- b) Have the identical font size,
- c) Have the same base y coordinate, i.e. they share a baseline, and
- d) The space between the two adjacent edges of any adjacent pair is in threshold class 0. This yields single characters or collections of characters, which form words or numbers.

To linearize the two dimensional layout of the characters into a 1-dimensional version using guidelines of mathematical expression arrangement. The rules are based on an original set specified by Anderson [21]. The grammar comprises 12 rules to deal with the dissimilar spatial relationships between sets of symbols including scripts, fractions, limits, enclosed symbols, matrices, cases, accents and symbols spread over several lines. The partition of text and math lines is grounded both on spatial positioning of the line with respect to the left and right margin of the page as well as the number of words on that line, where a word is a series of alpha characters, clustered by the preceding parsing step. A line is then used as a text line if it a) Comprises only a sequence of words, b) Comprises at least two successive words and the number of other expressions is not greater than the number of words, c) Comprises more than three successive words irrespective of the number of other expressions. Everything else will be treated as style mathematics.

The one dimensional linearized lines are parsed using a LALR parser, resulting in a parse tree that is used as a midway representation for succeeding interpretation into many output formats. Structural information in the parse tree can be exploited by these interpretation modules.

**Characters** The entire number of lines and PDF characters extracted from the PDF file. This can contain typical characters such as a, 1 or =, and characters which form part of a larger symbol. Several symbols, especially large ones, are made from various characters and lines.

**Symbols** The number of symbols recognized, after plotting characters to glyphs. This is often less than the number of characters extracted, due to multi-character symbols.

**Misrecognized** The number of symbols that cannot be transformed to Unicode. They occur when character names are improper or lost from the font encoding of the PDF.

**Missing** The number of orphan characters left over after glyph matching. No character recognition errors arose using Maxtract.

With an appropriate PDF file, Maxtract yields faultless character identification outcomes and can create a high class restoration of text and formulae in LATEX, with parse trees of formulae with semantic information.

### 2.2.3 Detection and Segmentation of Table of Contents [7, 8, 22]

To excerpt the structural information from the table of contents (TOC) to assist to make digital document library by identifying/segmenting the TOC page is done by [7], [8], and [22]. They [7] present fully spontaneous identification and segmentation of table of contents (TOC) page from scanned document. Table of contents (TOC) detection from scanned document pages is significant for a user of the digital document library as a directory for the contents of the books, journals, and reports etc.

The TOC is text lines with an organized format. Identification of TOCs' are based on discovering page numbers related with the name of sections, sub-sections or articles/author. The page number is considered as word as a whole as the rightmost word of a text line [7].

However for TOC-I the right aligned page numbers will have their echo in the vertical projection in the form of a secluded narrow hump at the rightmost section. Discovery is done by examine the difference in the number of characters in the rightmost word of each line in TOC-II. Fig.2.1 and Fig.2.2 showed the decision tree of the method [7]. This [7] technique is based simply on the spatial distribution of the associated modules and low resolution image may not critically disturb the enactment of the segmentation. They have verified this by a 4 fold reduction of the resolution (300 dpi to 75 dpi) of the input images and got a 7 fold timing enhancement with a lowest squalor in segmentation enactment [7].

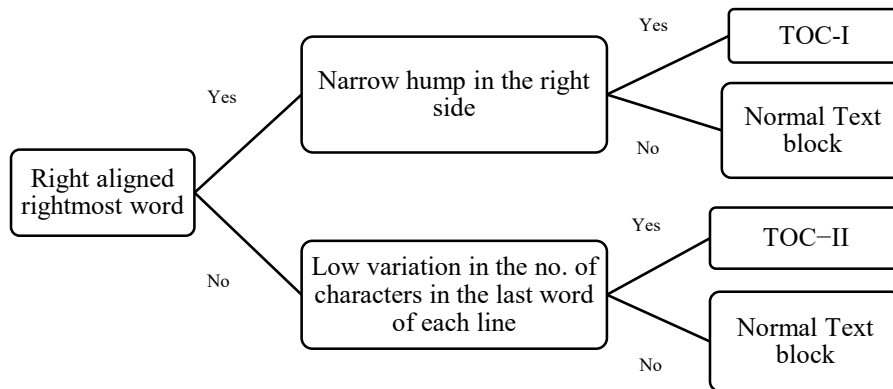


Fig 2.1: Decision Tree to Identify Two Types of Content Pages: TOC-I and TOC-II.

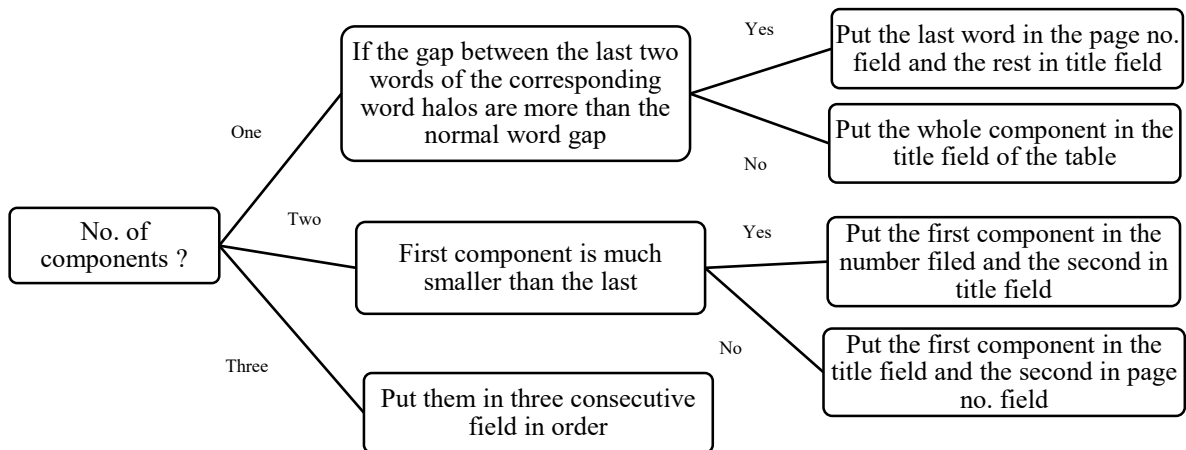


Figure 2.2: Decision tree for TOC segmentation

Moreover, in [8], the ToC identification is based on the subsequent rules: 1) a ToC is usually in the first few pages of the document, 2) a ToC typically contains some regularities of numbering and indentation, 3) a ToC commonly contains ordered references correlated to titles or sections in body pages. The last property can also be broken into 4 sub-properties: 1) Contiguity: a ToC consists of a series of contiguous references to some other parts, 2) Ordering: the references and the referred parts appear in the same order in the document, 3) No self-reference: all references refer outside the contiguous list of references, 4) Distinctness: the link from the references of ToC to the outside parts is injective, or every reference refers to a distinctive part.

Their method [8] does not rely on visual features such as font size or layout so that they could perform the detection purely based text, which is much more efficient for large scale extraction.

#### 2.2.4 Extracting Metadata Information [8]

In [8], they proposed a hybrid approach for extracting title and authors from a book that combines results from CiteSeer, a rule based extractor, and a SVM based extractor, leveraging web knowledge. For “table of contents” recognition, they proposed rules based on multiple regularities based on numbering and ordering. Furthermore, they also studied bibliography extraction and citation parsing for a large dataset of books. Lastly, they used the multiple fields accessible in books to rank books in answer to search queries. The system can successfully extract metadata and contents from great collections of online books and offers proficient book search and retrieval facilities. The metadata includes title, authors, ISBN, publish date and copyright. Since ISBN, date, copyright can be detected using strong rules, their main attention is on title and authors extraction. They developed innovative title and authors extractors based on experiential rules resultant from a small sample of books. They presumed that title and authors are constantly on the same page, i.e. the title page, and the title page is before ToC, foreword or preface which is shown in Table 2.1 and Table 2.24.

**Table 2.1.** Features Used in Book Metadata Extraction

<i>Feature</i>	<i>Description</i>
font size	Initial Font: the font size of the starting character Average Font: the average font size of all the characters Font Changes: number of changes in font size
location	Start X, End X, Start Y, End Y : the coordinates of the line block in the page Line Number: the (order) number of the line within the page, e.g. 2 indicates the second line Page Number: the (order) number of the page
text	Bag-of-words: Top 200 words selected by DF rank in the whole dataset; 1 indicates a word is in the line
others	Number of Words: the total number of words in the line Number of Digits: the total number of digital words in the line

#### 2.2.5 Extracting Bibliography [8]

Bibliography typically has obvious indicators such as “References”, “Bibliography” or “Sources”. However, unlike papers, books may have a bibliography at the end of each chapter. Thus, they examined bibliography in the entire body of book rather than in only the last few pages. If they recognized a line contains only one of the three keywords and

**Table 2.2.** Rules for Generating Venue Alias

<i>Rule</i>	<i>Examples of Venue Name</i>
None	IEEE Transactions on Pattern Analysis and Machine Intelligence
Transactions->Trans. Journal->J IEEE Trans. on Pattern Analysis and Machine Intelligence Proceedings->Proc	IEEE Trans. on Pattern Analysis and Machine Intelligence
Remove "of", "on", IEEE Trans. Pattern Analysis and Machine Intelligence "in", "the"	IEEE Trans. Pattern Analysis and Machine Intelligence
Acronymization	IEEE Trans. PAMI
Pure acronymization	PAMI
Manual edit	IEEE Trans. Pattern Anal. Machine Intell.

lines followed are ordered reference items, then they recognized it as a bibliography block. They explore the ordered number at the beginning of individual reference until there are no continuously increasing number found in the following 30 lines. 30 seems like a big distance for references. But they do find some references contained near 10 lines. Also they assumed that the distance between two bibliography blocks in two chapters will be much greater than 30.

### 2.2.6 Extraction of Data Points and Text Blocks [10]

In [10], they outline how data and text can be extracted automatically from these 2-D plots, thus removing a time consuming physical procedure. Their information extraction algorithm recognizes the axes of the figures, extracts text blocks like axes-labels and legends and finds data points in the figure. It also extracts the units appearing in the axes labels and segments the legends to identify the dissimilar lines in the legend, the different symbols and their related text justifications. Their algorithm also achieves the challenging task of splitting out overlapping text and data points successfully. They presented a suite of image analysis and machine learning algorithms that extract data and metadata related to it from the figures and its captions. The tool based upon these algorithms extracts data from the 2-D plots and keeps them in databases, so that this significant origin of data on the web can be searched using search engines. Precisely, the tool extracts the X and Y axis lines, ranges of values in the X and Y axes, the labels, units, and ticks on the axes, data points and lines in the plot, legends and the dissimilar types of data stated in the legend.

### 2.2.7 Summarizing Figures, Tables, and Algorithms in Scientific Publications [11]

Extracting a synopsis for a document-element from a digital document includes purifying information related to the document-element from the rest of the document. Resolving this problem precisely is easy if the semantics of the text is understood spontaneously. Nevertheless, state-of-the-art techniques of natural language processing and statistical text processing still fall short in fully understanding the semantics of text documents. Furthermore, good synopsis generation includes making a decision call concerning the level of detail that may be beneficial to an end-user. If a very huge synopsis is generated, it will be comprehensive, but the users' needs of finding information quickly will not be met. Their work is explained as-

- a) A method for extracting document-element related information from digital documents automatically. They treated the problem as a special case of query-biased summarization where the document-element itself is the query.
- b) A simple model for sentence selection that tries to strike a balance between the information content and the length of the synopsis. The top-ranked sentences selected by this model are lastly included in the synopsis.

The process followed by the [11], have three major parts which are as follows-

**1. Pre-processing:** The Steps are described in the following-

**a) Text Extraction:** They tried several tools available for PDF to text conversion (PDFBox [23], PDFTextStream [24], XPDF [25] and TET [26]) and they used PDFTextStream for their work. PDFTextStream preserves the sequence of text streams in the order they appeared in the document and for documents in double column format that are common in scientific literature.

**b) Document-Element Caption Parsing:** The CAPTION has 4 sub-parts. DOC\_EL\_TYPE specifies the type of the document element, namely figure, table or algorithm. FIG\_TYPE, TABLE\_TYPE and ALGO\_TYPE refer to the differences of the words "Figure", "Table" and "Algorithm" correspondingly, as they occur in the captions. The DOC\_EL\_TYPE non-terminal is followed by an integer that characterizes the document-element number. The integer is followed by a DELIMITER that can again be either ":" or ".". The last non-terminal TEXT gives a textual explanation of the element.



Identifying a grammar enables to follow a cohesive method for dealing with dissimilar types of document-elements. Figure 2.5 showed the grammar.

**Table 2.3.** A Grammar for Document-Element Captions

CAPTION	DOC_EL_TYPE   Integer   DELIMITER   TEXT
DOC_EL_TYPE	FIG_TYPE   TABLE_TYPE   ALGO_TYPE
FIG_TYPE	FIGURE   Figure   FIG.   Fig.
TABLE_TYPE	TABLE   Table
ALGO_TYPE	Algorithm   algorithm   Algo.   algo.
DELIMITER	
TEXT	A String of Characters

**c) Sentence Segmentation:** After extracting the caption sentences from the document text, they fragmented the document text into its constituent sentences. Their aim is to identify and extract sentences that are related to document-elements, correct sentence segmentation is very significant in this case. They have considered the average line length and word density.

**d) Reference Sentence Parsing:** To identify reference sentences, they used a grammar alike to that used for caption parsing. In the reference sentence, the delimiter will not be present in maximum cases and the integer will tell which element this sentence is referring.

**2. Feature Extraction:** The features are discussed below:

**a) Content based Features:** This feature utilizes information cues present in the caption. It is a score assigned to each sentence based on its similarity with the caption. Like captions, the reference sentences also contain important cues providing information about the document-elements. There are certain cue words and phrases that are used frequently by authors while describing a document-element.

**b) Context based features:** It is a binary feature with a value of 1 if a sentence is a reference sentence for the document-element. Otherwise, it has value 0. It is again a binary feature and has a value 1 if a sentence belongs to the same paragraph as the reference sentence. Otherwise, the value is 0. This feature captures the fact that a sentence closer to the reference sentence has a higher probability of being related to the document-element than a sentence located far away from the reference sentence.

**3. Classification:** The classification methods used by them for identifying document-element related sentences are a) Naive-Bayes Classifier and b) Support Vector Machines.

### **2.2.8 Extracting Algorithms in Scholarly Big Data [12]**

They proposed a set of amalgam procedures based on ensemble machine learning to discover pseudo-codes (PCs) and algorithmic procedures (APs) in scholarly documents. Precisely, three variations of a procedure for detecting PCs include an extension of the existing rule based method proposed by Bhatia, et al. [27], one based on ensemble machine learning techniques, and a hybrid of these two. The methods for discovering APs include a rule based method and a machine learning based method.

#### **1. Extracting Algorithms for PCs:**

**a) Rule Based Method:** A PC caption must contain at least one algorithm keyword, namely pseudo-code, algorithm, and procedure. Captions in which the algorithm keywords appear after prepositions are excluded, as these are not likely captions of PCs.

**b) Machine Learning Based Method:** These features are classified into 4 groups: Font-style based (FS), Context based (CX), Content based (CN), and Structure based (ST).

**c) Combined Method:** They proposed a combined method rule based method and the machine learning based method.

#### **2. Extracting Algorithms for Aps:**

**a) Rule Based Method:** AP indication sentences exhibit certain common properties: The sentences usually end with follows:, steps:, algorithm:, follows:, following:, follows., steps:, below:. And the sentences usually contain at least an algorithm keyword.

**b) Machine Learning Based Method:** The features can be categorized into two groups: Content based features (CN) and Context based features (CX).

Using these features and classifiers they have extracted the algorithms from the scholarly Data [12].

### **2.3 The Realization of Extracting Bangla Text, Image, Number and Knowledge**

Several work have been done on Bangla knowledge extraction. Some of the related methods are described in the following section.

### **2.3.1 Bangla Number Extraction and Recognition from Document Image [15]**

The method [11] can extract and identify Bangla numbers from a document image. This structure processes the image line by line of the text document. For each text line, the maximum and minimum widths of Bangla digits are assessed. Associated component labeling is used to filter the characters of these varieties of widths. The width cleaning outputs the Bangla digits if any and some single characters. The feature vector is extracted for each character of the filter output and fed the vector to the input of Multi-Layer Perception (MLP) for identification. Back-Propagation algorithm is used to train the Neural Network to recognize the digits only with 96% accuracy. The network has also the ability to find the character if it is not a digit.

To increase the classification efficacy, individual identification engine can be used to recognize letters, digits and other symbols. They composed the numbers (series of digits) by filtering out the word from the document image and to recognize in separate digit to produce the number as text.

### **2.3.2 Phrase-level Polarity Identification for Bangla [28]**

In this work [28], opinion polarity classification on news texts has been conceded out for Bangla language using Support Vector Machine (SVM). The scheme recognizes semantic direction of an opinionated phrase as either positive or negative. The cataloging of text as either subjective or objective is obviously a precursor to determining the opinion orientation of evaluative text since objective text is not evaluative by description. A rule based subjectivity classifier has been used. This system is a hybrid approach to the problem, works with lexicon entities and linguistic syntactic feature. They proposed a comprehensive opinion mining system that can recognize subjective sentences within a document and a proficient feature based automatic opinion polarity detection algorithm to identify polarity of phrases.

Bangla corpus attainment is a vital task for any NLP system development. For this task Bangla news corpus has been identified. News text can be divided into two main types: (1) News reports that aim to objectively present factual information, and (2) Opinionated articles that clearly present authors' and readers' views, evaluation or judgment about some specific events or persons.

In order to identify features we started with Part Of Speech (POS) categories and continued the exploration with the other features like chunk, functional word, SentiWordNet in Bangla [29], stemming cluster, Negative word list and Dependency tree feature. The feature extraction pattern for any Machine Learning task is crucial since proper identification of the entire features directly affect the performance of the system. Functional word, SentiWordNet (Bangla) and Negative word list is completely dictionary based.

### **2.3.3 Bangla Text Extraction from Natural Scene Images [30]**

In [30], they proposed scheme based on analysis of associated components for extraction of Devanagari and Bangla texts from camera captured scene images. A common feature of these two scripts is the presence of headline and the proposed scheme uses mathematical morphology processes for their extraction. As well, they consider some principles for robust sifting of text components from such scene images. They tested the algorithm on a repository of 100 scene images containing texts of Devanagari and or Bangla. A global binarization system like the well-known Otsu's method is typically not suitable for camera captured images since the gray-value histogram of such an image is not bi-modal. Binarization of such an image using a threshold value often leads to loss of textual information contrary to the background.

### **2.3.4 Sentiment Analysis on Bangla and Romanized Bangla Text [31]**

In this study [31] a significant textual dataset of both Bangla and Romanized Bangla texts have been delivered which is first of this kind and post-processed, several authorized, and ready for SA implementation and experiments. Additional, this dataset have been tested in Deep Recurrent model, exactly, Long Short Term Memory (LSTM), using two types of loss functions – binary cross-entropy and categorical cross-entropy, and also some investigational pre-training were directed by using data from one authentication to pre-train the other and vice versa. They explored –

- a) A Data set of 10,000 Bangla and Romanized Bangla text examples, where each sample was explained by two adult Bangla speakers
- b) Pre-processing the data in a way so that it is readily usable by researchers.
- c) Application of deep recurrent models on the Bangla and Romanized Bangla text corpus.

- d) Pre-train dataset of one label for another (and vice versa) to see if it gives better results.

The dataset contains of three groups –Positive, Negative, and Ambiguous. Data were composed from numerous micro-blog sites, such as, Facebook, Twitter, YouTube and some online news portal, product review panels etc. The model is based on Recurrent Neural Networks (RNN) more exactly they used LSTM neural network and also used Keras’ model-level library since it has all the important features to develop the deep learning model.

### **2.3.5 Bangla Text Summarization by Sentence Extraction [16]**

In this work [16], they have followed a meek and easy-to-implement method to Bangla single document text summarization since the refined summarization scheme involves resources for deeper semantic analysis. They have explored the impact of thematic term feature and position feature on Bangla text summarization. They have compared the proposed method to the LEAD baseline which was defined for single document text summarization task. LEAD baseline considers the first  $n$  words of an input article as a summary, where  $n$  is a predefined summary length. They used a lightweight stemmer for Bengali that strips the suffixes using a predefined suffix list, on a “longest match” basis, using the algorithm similar to that for Hindi [32]. After an input document is constructed and stemmed, the document is broken into a collection of sentences and the sentences are ranked based on the following features: Thematic term, Positional Value, Sentence length, Combining Parameters for Sentence Ranking.

### **2.4 The Realization of Extracting Keywords [33]**

Document-oriented methods therefore provide context free document features, enabling extra analytic methods such as those described in [34] and [35] that describe variations within a text stream over period. These document-oriented methods are suited to corpora that change, such as pools of published technical abstracts that grow over time or streams of news articles. Moreover, by working on a single document, these approaches intrinsically scale to massive collections and can be applied in several contexts to improve IR systems and investigation tools.

Previous work on document-oriented methods of keyword extraction has joint natural language processing methods to recognize part-of-speech (POS) tags that are combined with supervised learning, machine-learning algorithms, or statistical tactics.

#### **2.4.1 Rapid Automatic Keyword Extraction [33, 36, 37]**

Rapid Automatic Keyword Extraction (RAKE), an unsupervised, domain-independent, and language-independent technique for extracting keywords from separate documents. They delivered details of the algorithm and its configuration parameters, and present results on a benchmark dataset of technical abstracts, showing that RAKE is more computationally proficient than TextRank while accomplishing greater precision and analogous recall scores. They also described a new method for producing stoplists, which is used to configure RAKE for particular domains and corpora. They implemented RAKE to a corpus of news articles and describe metrics for evaluating the distinctiveness, essentiality, and generality of extracted keywords, allowing a system to recognize keywords that are vital or general to documents in the nonexistence of manual annotations.

RAKE is grounded on commonly comprise multiple words but rarely contain standard punctuation or stop words, such as the function words *and*, *the*, and *of*, or other words with minimal lexical meaning. The input restrictions for RAKE comprise a list of stop words (or stoplist), a set of phrase delimiters, and a set of word delimiters. RAKE uses stop words and phrase delimiters to divide the document text into candidate keywords, which are series of content words as they occur in the text. Co-occurrences of words within these candidate keywords are significant and permitted to detect word co-occurrence without the application of a randomly sized sliding window. Word relations are thus measured in a manner that spontaneously adjusts to the style and content of the text, enabling adaptive and fine-grained measurement of word co-occurrences that will be used to score aspirant keywords.

They assessed several metrics for calculating word scores, based on the degree and occurrence of word vertices in the graph: 1) Word frequency, 2) Word degree and 3) Ratio of degree to frequency

They followed the approach described in [36] using the testing set for assessment because RAKE does not require a training set. The Multilingual Rapid Automatic Keyword Extraction (mRake) is another version of Rake by [37] where the feature are-Automatic keyword extraction from text written in any language, No need to know language of text

beforehand, No need to have list of stopwords and 26 languages are currently available, for the rest - stopwords are generated from provided text

#### **2.4.2 Other Schemes [38, 39]**

Several schemes have also been examined in the field keyword handling. Such as Term frequency (TF) and keyword adjacency (KA), TextRank [36], TAKE [38], Swiftrank [39] etc. We have used TF, TextRank and mRake in our system for keyword generation so that we can get all frequently used words from the text.

### **2.5 Discussion**

All the models presented in this chapter have some pros and cons. However, for Bangla official documents there is no research have been done. In this circumstances, we propose a knowledge extraction model which will solve our problem. We also provide a natural language based model to find a knowledge. The detail of the proposed structure is presented in the fourth chapter after discussing the theoretical consideration in the next chapter.

## CHAPTER III

### Theoretical Consideration

#### 3.1 Introduction

The theoretical analysis is reviewed in this chapter. Some of the required knowledge to implement the system to get the result are presented here. By using the connection of the parts of the data we can extract the related information and knowledge [40] which is depicted in the Fig.3.1. Several techniques are used to extract information from the text which are described in the following section.

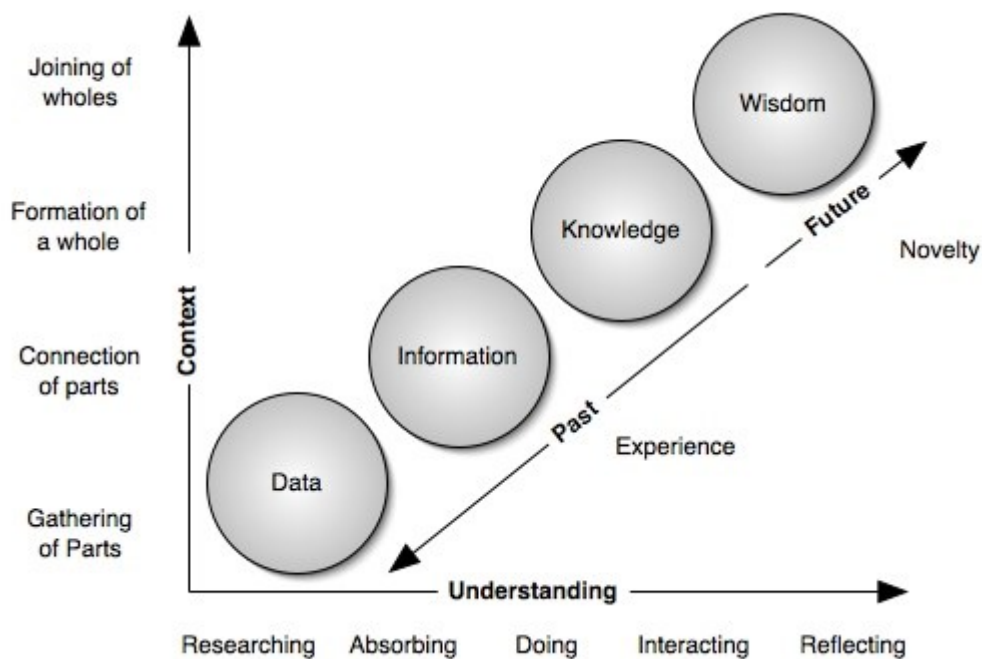


Fig.3.1: Data, Information, Knowledge, Wisdom Chain

#### 3.2 Word Density [11]

The document text comprises a lot of sparse lines equivalent to table data, equations, authors' names and affiliations etc. Normally, when altering from PDF to text, structure of table text etc. is mislaid and the mathematical symbols in equations are not correctly transformed to text form.



Hence, these need to be detached. In order to recognize these sparse lines, [11] used a word density measure that is stated as follows:

$$d_{wl} = \frac{L}{L + S} \quad (3.1)$$

Here,

$d_{wl}$  = the word density of line  $l$ ,

$L$  = the length of line  $l$  in words,

$S$  = the number of spaces in line  $l$ .

Note that the word density of a usual text line is larger than 0.5, because they [9] used,  $S = L - 1$ . They clean out only those lines from the document having word density less than 0.5. The cleaned up text is then sent to a sentence segmenter. It splits the document text into its principal sentences and produces the sentence set  $S$ .

### 3.3 Similarity with Caption [11]

In [9], they use information cues existing in the caption. It is a score assigned to each sentence grounded on its likeness with the caption. After elimination of stopwords from the caption sentence and stemming using Porter's Algorithm [41], the subsequent keywords form a "query" which offers cues about the information covered in the document-element is found. This query is then used to give Similarity Scores to all sentences in the document grounded on their similarity to the query. Okapi BM25 [42], [43] can be used as similarity measure, since it has been proved to be very beneficial in a wide diversity of IR jobs.

**Inverse Sentence Frequency:** It is similar in function to inverse document frequency (IDF) as used in information retrieval and minimizes common terms. The second term in Okapi BM25 [42] represents the frequency of individual query term in sentence, regulated by sentence length and scaled. Later calculating the scores for all the sentences, the highest ranked sentences with the top scores are picked.

### 3.4 Naive Bayes Classifier [11]

Naive Bayes classifiers have been formerly used fruitfully to extract sentences for document summarization [44], [45]. This process is meek, fast and can be comfortably adjusted for use in current digital libraries having lots of documents. It is described in the following.

Let the set of sentences that are connected to the document-element  $d$  be  $S_d$  and let  $S$  be the set of all sentences in the document  $D$ . Given the features  $F_1, F_2, \dots, F_n$  for sentence  $s \in S$ , the Bayes' rule to compute the probability that  $s$  also belongs to  $S_d$ , as follows:

If  $F_1, F_2, \dots, F_n$  are the presumed features for sentence  $s \in S$ , Bayes' rule calculates the probability that  $s$  belongs to  $S_d$ , as follows:

$$P(s \in S | F_1, F_2, \dots, F_n) = \frac{P(F_1, F_2, \dots, F_n | s \in S_d)P(s \in S_d)}{P(F_1, F_2, \dots, F_n)} \quad (3.2)$$

Where,

$S_d$  = set of sentences that are related to the document-element  $d$

$S$  = set of all sentences in the document  $D$ .

Assuming independent features, the above equation can be written as:

$$P(s \in S | F_1, F_2, \dots, F_n) = \frac{\prod_{i=1}^n P(F_i | s \in S_d)P(s \in S_d)}{\prod_{i=1}^n P(F_i)} \quad (3.3)$$

The probabilities  $P(F_i | s \in S_d)$  and  $P(F_i)$  are not identified a priori but they can be guessed by calculating frequencies in the training set. This gives a meek Bayesian classification function that gives a probability score to individual sentence in the document. The top-scoring sentences can be known as connected to document elements. The scores for all the sentences in the document are standardized in the range [0–1] in [9]. Here  $P(s \in S_d)$  is the identical for all sentences in the document and is therefore a constant. They are concerned in the relative values of sentence scores only and not the absolute values, so this constant is overlooked by [11].

### 3.5 Sentence Selection [11]

In overall, the sentence choosing problem can be outlined as follows by [9]: let  $U_k$  be the Utility measure of sentence  $S_k$  that expresses whether it is beneficial to select the sentence or not. It is explained as:

$$U_k = g(k) - f(k) \quad (3.4)$$

Here,  $g(k)$  is a function that favors the selection of  $S_k$  and  $f(k)$  is another function contrasting the selection of  $S_k$ . Sentences for which utility  $> 0$  are comprised in the final set. When  $g(k)$  the similarity is between  $S_k$  and the query and  $f(k)$  determines the redundancy of  $S_k$ ,  $U_k$  becomes similar as Maximum Marginal Relevance [9].

Let the score of the  $k^{th}$  sentence be  $score_k$  and let all sentences be graded in decreasing order of their scores so that  $i < j$  implies  $score_i \geq score_j$ . They [9] defined the Utility measure  $U_k$  in the following:

$$U_k = score_k - (1 - exp^{-\gamma(k-1)}) \quad (3.5)$$

### 3.6 Confusion Matrix [46, 47, 48, 49]

In the arena of machine learning and exactly in the statistical problem, a confusion matrix is known as an error matrix. It is a precise table layout that explains the conception of the performance of an algorithm. Each row of the matrix characterizes the occurrences in a predicted class while each column characterizes the occurrences in an actual class which is shown in Fig.4.1.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Fig. 3.2: Confusion Matrix

Where Condition positive (P) = the number of real positive cases in the data

Condition negative (N) = the number of real negative cases in the data

True positive (TP) = Equivalent with success

True negative (TN) = Equivalent with right refusal

False positive (FP) = Equivalent with wrong alarm, Type I error

False negative (FN) = Equivalent with failure, Type II error

Additionally, precision and recall is used to investigate a predictive model and compute these statistics over authentic or test dataset.

Precision is considered over the entire predictions of the model. It is the proportion between the correct predictions and the total predictions. In further words, precision specifies how worthy the model at whatever it predicted is.

$$\text{So } \textit{Pricision} = \frac{TP}{TP + FP} \quad (3.6)$$

However, *True Positive(TP) + False Positive(FP) = Total Predicted Positive*

*Total Predicted Positive* is shown in the Fig.3.3.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Fig.3.3: Confusion Matrix for Total Predicted Positive

$$\text{Now, } \textit{Pricision} = \frac{\textit{True positive}}{\textit{Total Predicted Positive}}$$

Recall is the Ratio of the correct predictions and the total number of correct items in the set. It is expressed as % of the total correct (positive) items correctly predicted by the model. In other words, recall indicates how good the model at picking the correct items is.

$$\text{Therefore, } \textit{Recall} = \frac{TP}{TP + FN} \quad (3.7)$$

Again, *True Positive (TP) + False Negative(FN) = Actual Positive*

*Actual Positive* is shown in the Fig.3.4

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Fig.3.4: Confusion Matrix for Actual Positive

$$\text{Now, Recall} = \frac{\text{True positive}}{\text{Actual Positive}}$$

Moreover, F1 Score is needed to poise between Precision and Recall and when there is an uneven class distribution (big amount of Actual Negatives). F1 Score is the weighted average of Precision and Recall. Therefore, this score takes together false positives and false negatives into explanation [49].

F1 is typically more beneficial than accuracy, particularly for rough class distribution. Accuracy works great if false positives and false negatives have analogous cost. If the cost of false positives and false negatives are very dissimilar, it is preferable to focus at both Precision and Recall.

$$F1 = \frac{2(\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \quad (3.8)$$

### 3.7 Graph-Based Centrality and PageRank and TextRank [36, 50, 51]

The elementary notion behind the PageRank [38] algorithm is that the prominence of a node within a graph can be decided by taking into account global information recursively calculated from the whole graph, with associates to high-scoring nodes contributing more to the score of a node than associates to low-scoring nodes. This significance can be used as a measure of centrality. PageRank gives to each node in a directed graph a numerical score between 0 and 1, known as its PageRank score ( $PR$ ), and defined as [36]

$$PR(V_i) = 1 - d + d \times \sum_{j \in InV_i} \frac{1}{|Out(V_j)|} PR(V_j) \quad (3.9)$$

Where

$InV_i$  = set of vertices that point to  $V_i$ ,

$Out(V_j)$  = set of vertices pointed to by  $V_j$ , and

$d$  = damping factor, typically set to around 0.8 to 0.9 [52]. Using the likeness of a random surfer, nodes visited more often will be those with numerous links coming in from other commonly visited nodes, and the role of  $d$  is to back up some probability for hopping to any node in the graph, thus stopping getting stuck in a disconnected part of the graph.

Though formerly suggested in the background of ranking webpages, PageRank can be used more commonly to decide the significance of an object in a network. For example, TextRank [36] and LexRank [53] utilize PageRank for ranking sentences for the purpose of extractive text summarization.

The underlying hypothesis for computing the prominence of a sentence is that sentences which are analogous to a big number of additional important sentences are central. Thus, by ranking sentences according to their centrality, the highest ranking sentences can then be mined for the summary.

In TextRank and LexRank, individual sentence in a document or documents is characterized by a node on a graph. Nevertheless, unlike a web graph, in which edges are unweighted, edges on a document graph are weighted with a value signifying the likeness between sentences. The PageRank can be simply adjusted to deal with weighted undirected edges, expressed as:

$$PR(V_i) = 1 - d + d \times \sum_{j=1}^N (w_{ji} \frac{PR(V_j)}{\sum_{k=1}^N w_{jk}}) \quad (3.10)$$

Where

$w_{ji}$  = similarity between  $V_j$  and  $V_i$  and

In [50], they assumed that these weights are kept in a matrix  $W = \{w_{ji}\}$ , which is referred as the “affinity matrix”. Here, the summations are over all objects in the graph. TextRank and LexRank put on a single instance of PageRank to the pool of sentences. A significant feature of TextRank is that it does not need profound linguistic knowledge, nor domain or language specific annotated corpora, which makes it extremely convenient to other domains, genres, or languages.

### 3.8 Mixture Models and the EM Algorithm [50]

The Mixture Model algorithm is exhibited as a linear combination of  $C$  component densities  $P(m|x)$  in the form  $\sum \pi_m P(m|x)$ , where the  $\pi_m$  are called mixing coefficients, and characterize the prior probability of data point  $x$  having been produced from

component  $m$  of the mixture. Supposing that the parameters of individual component are denoted by a parameter vector  $\theta_m$ , the problem is to determine the values of the components of this vector, and this can be accomplished using the Expectation-Maximization algorithm [42]. Succeeding random initialization of the parameter vectors  $\theta_m$ ,  $m=1, \dots, C$ , an Expectation step (E-step), followed by a Maximization step (M-step), are repeated until convergence. The E-step calculates the cluster membership probabilities. For example, supposing spherical Gaussian mixture components, these probabilities are calculated as:

$$P(m|x_i) = \frac{\pi_m P(x_i|\mu_m, \sigma_m)}{\sum_{k=1..C} \pi_k P(x_i|\mu_k, \sigma_k)}, \quad m=1, \dots, C, \quad (3.11)$$

where  $\mu_m$ , and  $\sigma_m$  are the present guesses of the mean and standard deviation, respectively, of component  $m$ . The denominator acts as a normalization factor, ensuring that  $0 \leq P(m|x_i) \leq 1$  and  $\sum_{m=1}^C P(m|x_i) = 1$ . In the M-step, these probabilities are then used to re estimate the parameters. Over using the spherical Gaussian case-

$$\mu_m = \frac{\sum_{i=1}^N P(m|x_i) x_i}{\sum_{i=1}^N P(m|x_i)} \quad m=1, \dots, C, \quad (3.12)$$

$$\delta_m^2 = \frac{\sum_{i=1}^N P(m|x_i) \|x_i - \mu_m\|^2}{\sum_{i=1}^N P(m|x_i)}, \quad m=1, 2, \dots, C, \quad (3.13)$$

$$\pi_m = \frac{1}{N} \sum_{i=1}^N P(m|x_i), \quad m=1, 2, \dots, C, \quad (3.14)$$

$P(x|\mu, \delta)$  are called ‘‘likelihoods’’ and in the case of Gaussians are just the value of the Gaussian with mean and variance  $\delta^2$  calculated at point  $x$ .

### 3.9 Discussion

In this chapter we present the theoretical analyses for the proposed structures. Without understanding the theory of the discussed models above it is very hard to understand our system which is described in the following section.

## CHAPTER IV

### Methodology

#### 4.1 Introduction

The amount of official Bangla files are produced by professionals, Judicial, and academician and is very difficult to find the previous or specific information. Therefore, our proposed system correctly detect the knowledge and show the result to the user with the following outcomes are: It extracts Information for Bangla legal documents where the user will find his/her desired information using a set of the various knowledge-based method in Bangla text. Automatically discover and extract the decisions and agendas from official documents and make an analysis and classification for a sample Dataset to detect knowledge.

#### 4.2 Realization of the Method for Proposed Bangla Knowledge Extraction

In this article, the process of extracting decisions, agenda and the query result of the user with keywords form Bangla PDFs are presented. Fig 4.1 demonstrates the high-level design of the recommended system. Firstly, the Bangla documents are processed to identify the agenda with the decision. Then, the extracted knowledge provides the date and the meeting number from the document which is presented chronologically. The algorithm for knowledge extraction is given the presented in Fig 4.2. In the algorithm there are two algorithms which are *Extraction ()* and *Decision Extraction with features ()*. *Extraction* algorithm extracts all decisions and stored in the decision pool. From the decision pool the user query is found out by calling *Decision Extraction with features* algorithm.

To identify the desired lines, we processed the document to get the pure text and then with the corresponding features we extracted the knowledge from the text and arrange the extracted information with the meeting date and meeting number. From these extracted decisions we extracted the keywords using different algorithm then form these data we have selected the most frequent keywords from the various algorithm. These keyword maintain a knowledge base with synonyms from the domain and mapped with user query. Then the user findings are shown as a result.



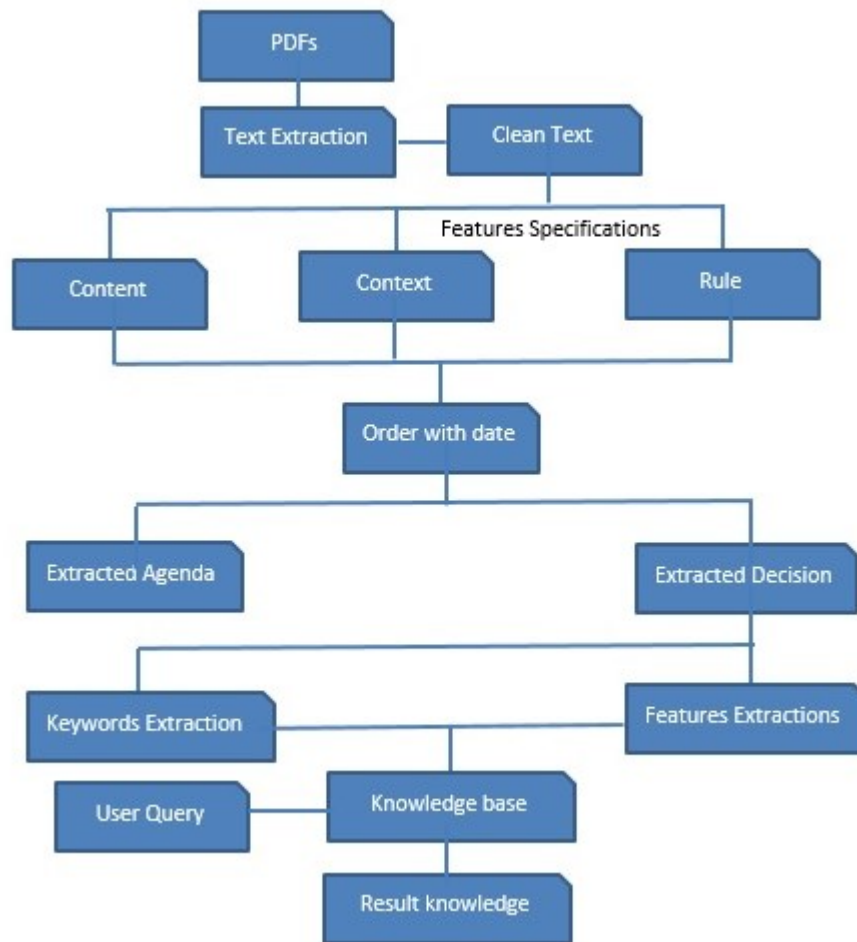


Fig. 4.1: Design of Proposed Knowledge Extraction from Official Bangla Documents

We have divided the task into major three parts as extraction of agenda and decisions text, ordering the documents chronologically and finding user query from the extracted decision pool. The algorithm of knowledge extraction from official Bangla documents is given below. Algorithm extraction () and Decision Extraction with features () are given in Fig.4.2 and Fig.4.4

*Algorithm: **Knowledge Extraction** ()*

1. For all pdf files call algorithm **Extraction** () and print result
2. Decision pool = Make a set of decision sentence
3. From Decision pool call algorithm **Decision Extraction with features** ()
3. Print result with relevant information from Step 3
4. Classify all decision sentences.

Fig. 4.2: Proposed Knowledge Extraction Algorithm

The design for Extracting agenda and decisions and Algorithm for Extracting agenda and decisions are shown Fig 4.3 and Fig.4.4 respectively. For each documents the system extracts the required sentences are using different features described below. The system takes the PDF documents  $D_i$ , a set of a Bangla documents as input and a Query  $Q$  and returns  $d_i$ , a set of Bangla documents with the detected results respectively. The major parts of the proposed model are given in the following subsection.

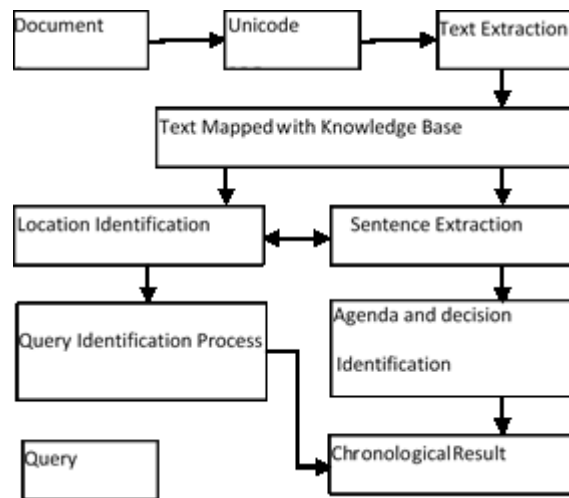


Fig. 4.3: Design of Agenda and Decision Extraction from Official Bangla Documents

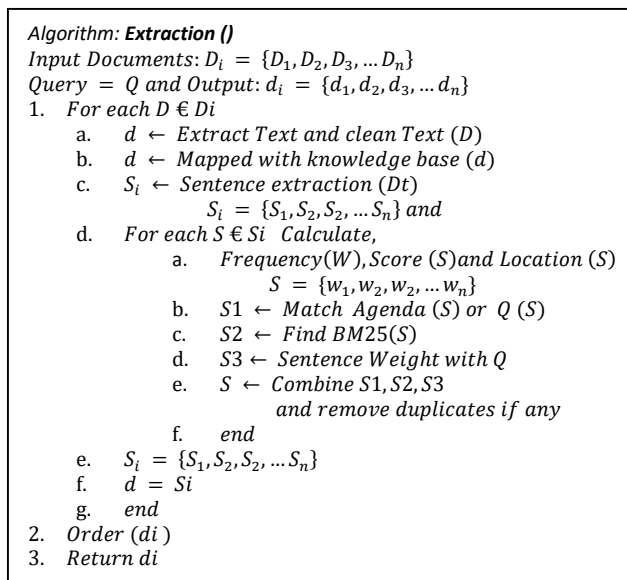


Fig. 4.4: Proposed Algorithm for Agenda and Decision Extraction from Bangla Documents

Fig 4.5. Shows the system for keyword extraction and finding user query from the documents and Fig.4.6. Shows the algorithm respectively. The system takes the documents  $De_i$ , a set of Bangla decision and Query  $Q$  and returns  $de_i$ , a set of Bangla detected texts. For each query, the system collects the required lines with some semantics features explained in the following sub sections.

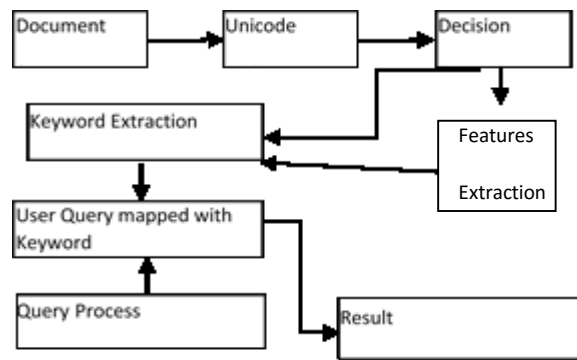


Fig. 4.5: Design of Keyword Extraction and Finding User Query from Documents

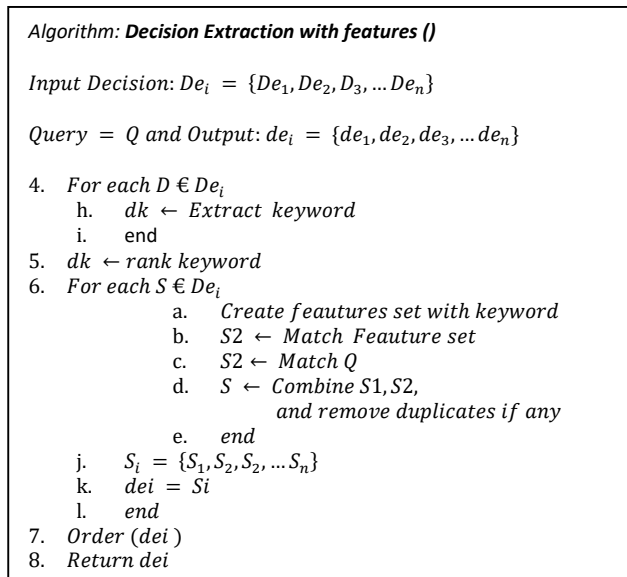


Fig. 4.6: Proposed Algorithm for User Query Extraction with Features from Decision Pool

### 4.3 Data Selection and Pre-processing

The system takes the PDF documents  $D_i$ , a set of a Bangla documents as input and a Query  $Q$  and returns  $d_i$ , a set of Bangla documents with the detected results respectively.

### 4.3.1 Selection of the Target Data

For the selection of the data, we have collected the resolutions of the academic council of Khulna University of Engineering & Technology (KUET) as our target data for a case study where data are written in the Bangla doc files. These files should be written in a Unicode format. Otherwise the PDF files cannot be read by our process. The Unicode font must be embedded with the doc file before making the PDF format. These files have specific format with structured data where different agendas with its decisions are presented here from the meeting of academic council of KUET. We have experimented with 29 resolutions of Academic Council's meeting of KUET for agenda and decision extraction and for user query extraction.

### 4.3.2 Pre-processing the Data

Inspired by Hassan [55] and Bhatia, et al., [12], we used PDFBox [23] to extract text and then cleaned the sentences like removing tables, web and email address and, references. Here the PDF should be made from a Unicode file and the font should be embedded with the file. After pre-processing the data is transformed to a specific format from where we can mine the required data.

- 1) For Bangla vowel marks, there are some broken words in the text. For this problem, we have made our own knowledge base for this specific domain. Such as “প্রকণতা” will be “প্রণেতা”. Broken words from the document then, searched and matched with our knowledge base words and the corresponding broken words with the correct one is then altered. We have made a knowledge base for 565 words from 12 documents from the target data. This works as a small database to retrieve the words from the files.
- 2) The PDFBox prints the sequence of text in the order they appeared in the document [14], [9]. Thus all the line in the text is given a sequence number for measuring the location of the sentences.
- 3) All the sentences are then extracted using stopword in Bangla (।) and Removing different punctuations, the word frequency is calculated.
- 4) We gave a score to all sentences using the following equation where the word or term frequency in a sentence is summed.

$$\text{Sentence Score} = \sum \text{Term Frequency in a sentence} \quad (4.1)$$

5) Then sentence weight is calculated with query value [56].

$$\text{Sentence Weight} = \text{Query term value} + \sum \text{Term Frequency in a sentence} \quad (4.2)$$

#### 4.4 Feature and Patterns Specifications for Decision Extraction

In the PDF file, there is a word “আলোচ্যসূচী” which means “agenda” and the point is discussed under these word with a number of the meeting and a number of the topics. There is a discussion immediate after the “agenda”. A decision may appear or not after the discussion and if appeared it is written as “সিদ্ধান্তঃ” with a vowel mark “:” or “Bisarga” which means “decision” and we will describe “আলোচ্যসূচী” as “agenda” and “সিদ্ধান্তঃ” as “decision” for the rest of the paper. The aim is to fetch the “agenda”, “decision” and the query words by the user, if it is discussed in the document.

For this, we have used three features where a phrase or keyword is matching directly within the words in sentences. However, with the direct search like this which we named as Rule-based features or exact phrase rule features, we are not able to find the exact information we are looking for [11], [12]. Hence we have used the other two features. The features are discussed below:

##### 4.4.1 Query-Based Features

1. The phrase “agenda” is directly searched for extracting the agenda in a line. A set of a regular expression is used in this process. All the lines which contain the phrase, are fetched by the regular expression (regex).

This is also true for decision extraction. So here we need content based features described later. However, in this means we only found out the words exactly matched the phrase and we did not get relevant information regarding our search. In this case, we have used the context based features for a better extraction.

**Example 4.1** in the Fig. 4.7 we find there are two agenda and by only using rule base features we will get only the first line from the text with agenda keyword which means will get only one line with agenda. It will not return the next list of the agenda and do not get the relevant information

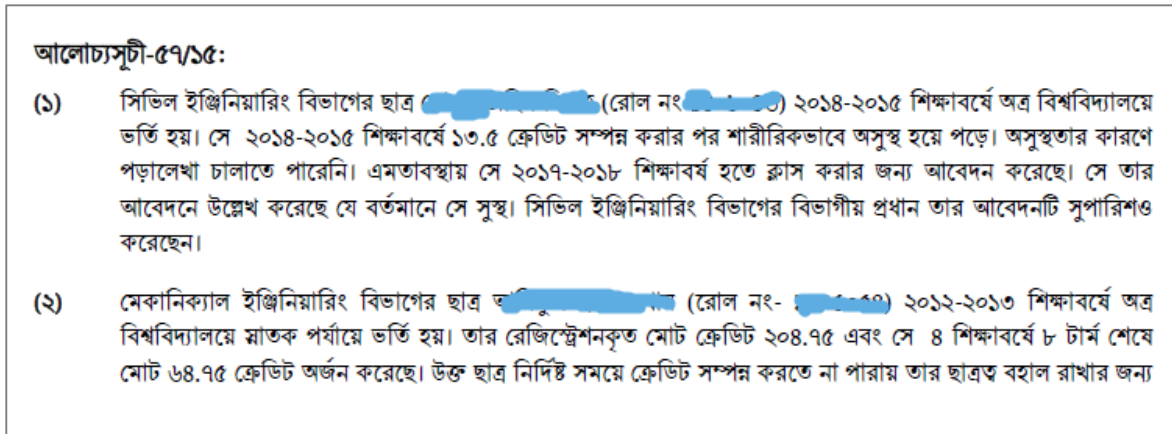


Fig. 4.7: Example 4.1

2. Similarly, “decision” phrase or query word are found using regex from all the documents. Moreover, Sometimes the author of the document can write the “সিদ্ধান্তঃ” as “সিদ্ধান্ত“. In this point only exact matching will not give us the perfect result, because a line may contain this phrase as reference but not as a decision. Nevertheless, there are many lines which contains “decision” of the “agenda” under the “decision” phrase and don’t contain the phrase. That’s why we need to search for content and context based features. And for the user query, the same above process is used to extract the information from the documents. Then the similarity of the extracted sentences are measured by content-based features.

3. Then, to calculate the similarity of the sentences with the query or “agenda” or “decision”, we compute the similarity scores to all sentences in the document based on their similarity to the query [11]. We have used Okapi BM25 [42], [43], [11] to know the similarity of the sentences, meanwhile, it is verified as a very effective procedure in various information retrieval method by expert [11]. The equation of Okapi BM25 is given below:

If Q is the user query then the BM25 score of sentence S in document D is calculated as:

$$BM25(Q, S) = \sum_{t \in Q} \left\{ \log \frac{N}{sf_t} \times \frac{(k_1+1)tf_{ts}}{k_1 \left( (1-b) + b \times \left( \frac{L_s}{L_{av}} \right) \right) + tf_{ts}} \times \frac{(k_3+1)tf_{tq}}{k_3 + tf_{tq}} \right\} \quad (4.3)$$

Where N = Total Number of sentences in a PDF

$sf_t$  = Number of sentences that contain the term t in query Q = 1

$tf_{ts}$  = Frequency of term t in sentence, S

$tf_{tq}$  = Frequency of term t in query, Q

$l_s$  = Length of the sentence, S

$l_{av}$  = Average length of a sentence in the Document

$k_1, k_3$  and,  $b$  are the constant values which is fixed to 2, 2 and 0.75 respectively [11]. The part of the (4.3), the term,  $\log \frac{N}{sf_t}$  is known as inverse sentence frequency [11]. In the experiments, we have used a single query term, and for this reason the, query term is 1 here. We used this to measure the similarity of the extracted sentences and also the similarity between equation (4.2) and (4.3) to settle the context of the sentence.

#### 4.4.2 Content-Based Features

We found there are some definite phrases that are used frequently while describing a “decision” on the “agenda”. We have listed 77 phrases analyzing 12 documents manually and trained our model with these phrases given in Fig.4.8. We noticed that these phrases are placed at the end of the line of a decision-making sentence. These are decision making lines referring to make some remarks such as to permit some request or to order to do something or making a change in the system etc. which means all the phrases are verbs which point out to do something in the future or approve/order/suggest for a particular instance.

However, we have tried to make a content-based feature with the single word or verb from the mentioned phrases, but there are many lines which contain a single word from these phrases in the discussion section of the documents which are not “decision” taking sentences. For this reason, we used a phrase instead of a single verb. For an example, the verb “হবে” is used in the many places in the documents not only in the “decision” line. So we used the phrase “বাতিল করা হবে”, “জমা দিতে হবে”, “প্রেরণ করতে হবে” etc. instead of only “হবে”.

Context Features	অনুমোদন করা হলো	কমিটি গঠন করা হলো	সিদ্ধান্ত স্থগিত থাকবে
সিদ্ধান্ত গৃহীত হলো	সুপারিশ করা হলো	ধাকার চনা করা হলো	অনুরোধ করা হলো
নির্ধারণ করা হলো	বহাল রাখা হলো	কার্যকর হবে	কার্যকরী হবে
বলা হোক	অর্জন করতে হবে	বিবেচিত হবে	দাখিল করবেন
প্রদান করা হোক	ক্ষমতা প্রদান করা হলো	মনোনয়ন প্রদান করা হলো	প্রদান করা হলো
ধন্যবাদ প্রদান করা হয়	ধন্যবাদ জানানো হলো	ধন্যবাদ আপন করা হলো	প্রেরণ করা হোক
প্রেরণ করতে পারবেন	প্রেরণ করতে হবে	ব্যবস্থা নিতে হবে	ব্যবস্থা নিবেন
অবহিত করতে হবে	ভর্তি হতে হবে	অনুমোদন করা গেল না	সম্পন্ন করতে হবে
প্রযোজ্য হবে	অব্যাহতি পাবে	অবহিত হলেন	অনুমোদন করা হবে
অবহিত করতে হবে	জমা দিতে হবে	অনুমতি দেয়া যেতে পারে	নিশ্চিত করতে হবে
বখিত করা হলো	বাটিল করা হবে	পরামর্শ দেয়া হয়	নিশ্চিত করা হলো
বৃদ্ধি করা হলো	বাটিল হয়ে যাবে	আদা হোক	সন্মত প্রকাশ করা হয়
নিরাসিত হতে হবে	বরাদ্দ পাবে না	প্রয়োজনীয় ব্যবস্থা গ্রহণ করবেন	পরামর্শ দেয়া গেল
ধাকা চলবে না	রেজিস্ট্রেশন করতে হবে	পুনঃ উপস্থাপন করা হোক	পরামর্শ দেয়া হলো
অভিযোগ থাকতে পারবে না	ধাকা চলবে না	ব্যবস্থা নেয়ার প্রয়োজন নেই	সেরং পাঠানো হলো
ব্যবস্থা নিতে হবে	সকলে মত প্রকাশ করে	অনুরোধ জানানো হলো	বিবেচনা করা সম্ভব হলো না
3.75 এর কম নয়	অংশগ্রহণ করতে পারবে	সকল সদস্য একমত হন	রিপোর্ট পেশ করবে
পরীক্ষা অন্তর্ভুক্ত হবে	অন্তর্ভুক্ত করা হলো	বছর থাকবে	প্রত্যাহার করা হলো
বিবেচনা করা যেতে পারে	সংযুক্ত হবে	অনুমতি দেয়া হলো	অংশগ্রহণ করতে পারবে না
সচেতন থাকার চনা করা হলো	সংশোধন করা হলো		

Fig. 4.8: Decision Making Phrases

#### 4.4.3 Context-Based Features

1. All of the above similarity and matching processes only consider the content, even the rule or phrase mapping method. However, some sentences cannot be found by these procedures. There are many segments with numbering system in the documents which are listed as “agenda” and “decision”.

In these cases, we need to identify the context of the sentences. To identify the nearby sentence which is contextually related with the content, we identified the location of the sentence from the sequence of the text.

Such as, we found some sentences which do not contain the phrase but they are actually “agenda” with relevant information or continuation of the “agenda” on a different level. In this case, we have found out the position of the word in the sentence and if it is not the first word of the sentence, it is removed from the “agenda” list. If there is a numbering system immediately after the “agenda” word in the beginning of the sentence then we looked for the rest of the numbering sequence as an “agenda” continuation.

However, the “agenda” listed as a number, cannot be fetched by rule-based or content based procedures.



**Example 4.2** The 3<sup>rd</sup> point with the agenda is not added to the list shown in the Fig. 4.9. We have used the position of the decision and agenda from the text to extract the knowledge. This also same for the decision extraction.

<p><u>আলোচ্যসূচী-৫৭/০৬:</u></p>	<p>অত্র বিশ্ববিদ্যালয়ের সিভিল ইঞ্জিনিয়ারিং অনুষদের নির্বাহী কমিটির ৩৩তম সভার সুপারিশ অনুমোদন বিবেচনা প্রসঙ্গে।</p>
<p>অত্র বিশ্ববিদ্যালয়ের সিভিল ইঞ্জিনিয়ারিং অনুষদের নির্বাহী কমিটির ৩৩তম সভা গত ২০/১২/২০১৭ইং তারিখে অনুষ্ঠিত হয়। উক্ত সভায় নিম্নলিখিত বিষয়ে সুপারিশ করা হয়েছে।</p>	
<p>(১)</p>	<p>সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগের বিভিন্ন বর্ষের পরীক্ষা কমিটিসমূহ একাডেমিক কাউন্সিলে প্রেরণের বিষয় বিবেচনা।</p>
<p><b>সিদ্ধান্তঃ</b> নির্বাহী কমিটিতে বিস্তারিত আলোচনা শেষে সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগসমূহের বিভিন্ন টার্ম পরীক্ষার পরীক্ষা কমিটিসমূহ একাডেমিক কাউন্সিলে অনুমোদনের জন্য প্রেরণের সিদ্ধান্ত গৃহীত হয়।</p>	
<p>(২)</p>	<p>সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগের একাডেমিক ক্যালেন্ডার একাডেমিক কাউন্সিলে প্রেরণের বিষয় বিবেচনা।</p>
<p><b>সিদ্ধান্তঃ</b> সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগের একাডেমিক ক্যালেন্ডার কিছু সংশোধনসহ একাডেমিক কাউন্সিলে অনুমোদনের জন্য প্রেরণের সিদ্ধান্ত গৃহীত হয়।</p>	
<p>(৩)</p>	<p>আর্কিটেকচার বিভাগের ACUG-এর সভার <u>আলোচ্যসূচি ও সিদ্ধান্তঃ ০৪/০৩</u>-এর আলোকে আর্কিটেকচার বিভাগের ২য় বর্ষ ১ম টার্মের প্রস্তাবিত কোর্স কারিকুলাম একাডেমিক কাউন্সিলে প্রেরণের বিষয় বিবেচনা।</p>

Fig. 4.9: Example 4.2

2. If the sentence having the “decision” is immediately before an “agenda” and have some numbering sequences after the “decision” sentence and contain the content-based features, then the sentence enlisted to the decision list. Nevertheless, there are many sentences which also contain “decision” as a suggestion to the authority but not the “decision” making sentence in our case.

Moreover, if the “agenda” has a numbering order, then the agenda would be more than one for a topic. So thus finding the location we have extracted the sentences. After extracting the numbering decision list then we checked the immediately preceding “decision” containing sentence whether it contains specific phrases: “নিম্নরূপ সিদ্ধান্ত গৃহীত হয়” or “নিম্নলিখিত সিদ্ধান্ত গৃহীত হয়” etc. If it is found there then it is enlisted as “decision” sentences. The extra sentences which do not preserve the above criteria we have excluded that from the list.

**Example 4.3** Here we can see that the position of decision is very important to get the information and it should be before the agenda as it is described above feature in Fig. 4.10.

(8) বিইসিএম বিভাগের ১ম, ২য়, ৩য় ও ৪র্থ বর্ষ-এর প্রশ্নপত্র প্রণেতা ও পরীক্ষক প্যানেল- ২০১৭ এ নতুন শিক্ষকদের (বহিরাগত) নাম অন্তর্ভুক্তি করণের জন্য একাডেমিক কাউন্সিলে প্রেরণের বিষয় বিবেচনা।

সিদ্ধান্তঃ সভায় বিস্তারিত আলোচনা শেষে বিইসিএম বিভাগের ১ম, ২য়, ৩য় ও ৪র্থ বর্ষ-এর প্রশ্নপত্র প্রণেতা ও পরীক্ষক প্যানেল- ২০১৭ এ নতুন শিক্ষকদের নাম একাডেমিক কাউন্সিলে অনুমোদনের জন্য প্রেরণের সিদ্ধান্ত গৃহীত হয়।

এক্ষনে সিভিল ইঞ্জিনিয়ারিং অনুষদের নির্বাহী কমিটির ৩৩তম সভার সুপারিশসমূহ অনুমোদনের জন্য সভায় পেশ করা হলে নিম্নরূপ সিদ্ধান্ত গৃহীত হয়।

**সিদ্ধান্তঃ** ✓

(১) সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগের বিভিন্ন বর্ষের পরীক্ষা কমিটিসমূহ অনুমোদন করা হলো।

(২) সিভিল ইঞ্জিনিয়ারিং অনুষদের বিভিন্ন বিভাগের একাডেমিক ক্যালেন্ডার অনুমোদন করা হলো।

(৩) আর্কিটেকচার বিভাগের ২য় বর্ষ ১ম টার্মের প্রস্তাবিত কোর্স কারিকুলাম অনুমোদন করা হলো।

(৪) বিইসিএম বিভাগের ১ম, ২য়, ৩য় ও ৪র্থ বর্ষ-এর প্রশ্নপত্র প্রণেতা ও পরীক্ষক প্যানেল- ২০১৭ এ নতুন শিক্ষকদের (বহিরাগত) নাম অন্তর্ভুক্তি করণের বিষয়টি অনুমোদন করা হলো।

**আলোচ্যসূচী-৫৭/০৭:** অত্র বিশ্ববিদ্যালয়ের ইলেক্ট্রিক্যাল এন্ড ইলেক্ট্রনিক ইঞ্জিনিয়ারিং অনুষদের নির্বাহী কমিটির ৩৫তম সভার সুপারিশ অনুমোদন বিবেচনা প্রসঙ্গে।

অত্র বিশ্ববিদ্যালয়ের ইলেক্ট্রিক্যাল এন্ড ইলেক্ট্রনিক ইঞ্জিনিয়ারিং অনুষদের নির্বাহী কমিটির ৩৫তম সভা গত ১৯/১২/২০১৭ইং তারিখে অনুষ্ঠিত হয়। উক্ত সভায় নিম্নলিখিত বিষয়ে সুপারিশ করা হয়েছে।

(১) অত্র অনুষদভুক্ত বিভিন্ন বিভাগ কর্তৃক প্রেরিত পরীক্ষা কমিটি ২০১৭ (টার্ম-২) অনুমোদনের সুপারিশ প্রসঙ্গে।

**সিদ্ধান্তঃ** X ইইই, সিএসই, ইসিই, বিএমই ও এমএসই বিভাগের প্রস্তাবিত বি.এসসি. ইঞ্জিনিয়ারিং এর বিভিন্ন বর্ষের টার্ম-২ পরীক্ষা কমিটি ২০১৭ অনুমোদনের জন্য একাডেমিক কাউন্সিলের সভায় প্রেরণ করার সিদ্ধান্ত গৃহীত হয়।

Fig. 4.10: Example 4.3

3. Additionally, there are many “ ” quotation marks with “decision” word in a line immediately before the actual “decision” sentences which actually suggests some decision to authority.

We exclude this type of “decision” sentences as they are not the ultimate decision taken by the authority. However, if there is a duplicate sentence then it is removed from the list.

**Example 4.4** The decision with the quotation marks is shown in the Fig. 4.11 which we have excluded from the list of decision.

গত ১০/০৮/২০১৭ইং তারিখে অত্র বিশ্ববিদ্যালয়ের একাডেমিক কাউন্সিলের ৫৬তম সভায় [redacted] বিভাগের M.Sc. Eng. কোর্সের [redacted] (রোল নং-[redacted]) এর ফলাফল ও ডিগ্রী অনুমোদনের জন্য উপস্থাপন করা হলে নিম্নরূপ সিদ্ধান্ত গৃহীত হয়।

X **সিদ্ধান্তঃ** অত্র বিশ্ববিদ্যালয়ের [redacted] বিভাগের [redacted] (রোল নং-[redacted]) এর ফলাফল পরবর্তী সভায় পেশ করা হোক। [redacted] বিভাগের মাস্টার্স ডিগ্রী সম্পন্ন করার জন্য কোন থিওরী কোর্স সম্পন্ন করা বাধ্যতামূলক কিনা তা যাচাই-বাছাই করে একটি রিপোর্ট তৈরী করার জন্য একটি কমিটি গঠন করা হোক।” X

উপরোক্ত সিদ্ধান্ত মোতাবেক বিশ্ববিদ্যালয় কর্তৃপক্ষ মেকানিক্যাল অনুষদের ডীনকে সভাপতি করে একটি কমিটি গঠন করে।

Fig. 4.11: Example 4.4

#### 4.5 Ordering the Documents Chronologically

All the files have a date at the beginning of the content with meeting number. We have extracted the date with the meeting number and named the file accordingly. The meeting number of the file is then sorted and presented with the extracted information of date and meeting number.

**Example 4.5** The chorological information is found from the text and stored with the file name which is shown in the Fig. 4.12. We have stored the date of the meeting.

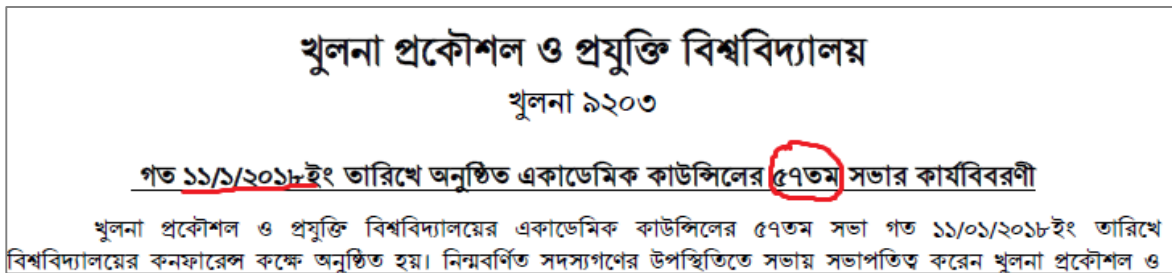


Fig. 4.12: Example 4.5

#### 4.6 Processing of the Keywords from the Extracted Decision Pool

In this section, the process of keywords extraction from the Bangla minutes of meeting of the academic council of KUET are presented. First of all from the processed Bangla documents we started to extract the keywords. We assumed the all the Bangla words are correctly extracted. Inspired by Bhatia [11], [12], the extracted set of decision is used and then the sentences are cleaned like removing stopwords. Here the domains specific stopwords are embedded with common stopwords. Then keywords extraction is done with different matured algorithms described as follows-

1. There are some common stop words in the text, Such as “অথবা”, “অনুযায়ী”, “অনেক”, “অনেকে”, “অন্তত”, “অন্য” etc. We have used 387 common Bangla stopwords and then listed 80 domain specific stopwords to the main stopwords list. Here are some example given in the Fig 4.13.

2. After stop words, elimination from the list of the decision, the term frequency-inverse document frequency, tf-idf [11] or TFIDF which reflect important word in a document or collection, is used to extract the keyword. Term frequency for the document:

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (4.4)$$

Where  $d$  is the document and

$t$ , is a term which means the number of times it occurs in  $d$  and  $\sum_{t' \in d} f_{t',d}$  is the total number of words in  $d$ .

The inverse document frequency is a measure of how much information the word provides [12].

$$idf(t, D) = \log \frac{N}{N^t} \quad (4.5)$$

Where  $N=|D|$  is the total number of documents and  $N^t$  is the number of documents where the term  $t$  appears.

And the  $tf-idf$  is [12]

$$tf-idf(t, d, D) = tf_{t,d} \cdot idf(t, D) \quad (4.6)$$

কর্ষক্রম	করেন	শিক্ষার্থী	অন্যান্য	দেওয়া	তিথি	কোর্স
সকলে	Eng	শিক্ষার্থীর	অনুষ্ঠিত	যাবে।	যেক।	কোর্সসমূহ
সকল	তিথি	২০১৭	সংক্রান্ত	পরিশ্রুত	যেক	অনুমোদন
সভায়	মোসা	ধন্যবাদ	১ম	চালিয়ে	উক্ত	অনুমোদনের
সভা	রয়েছে	বিভিন্ন	পরিশ্রুত	যাওয়ার	হবে।	গ্রহণের
ধৃত	নিম্নবর্ণিত	অনুশ্রুত	৩০/০৬/২০১৫	ইহা	পর্ষায়	সুপারিশ
উপলব্ধ	সদস্যবৃন্দ	অনুষ্ঠিত	১০০	হলো	পর্ষায়ের	বিভাগের
উপলব্ধ	শিক্ষার্থীকে	২০১৬-২০১৭	জ্ঞাপন	হলো।	in	বিভাগ

Fig. 4.13: Stopwords for the the Domain

3. A well-known natural language processing algorithm, Rapid Automatic Keyword Extraction (RAKE) can spontaneously extract keywords from documents [33]. We have used Multilingual Rapid Automatic Keyword Extraction (mRake) [37]. For the purposes to detect the keywords, if it used the entire documents in mRake its stopwords will be more specific. However, it is language independent. In this mRake the stopwords are generated from the text itself and the more text the more stopwords [37].

TextRank is described in the[36],[50],[52], creates a graph of the words and relationships between them from a document, then finds the most important points of the words based on importance scores calculated from the entire words graph. We have used this algorithm to rank the sentences here.

4. Then we have selected the most common 20 keywords from these techniques. The most frequently extracted keywords are listed with the  $tf-idf$  and rake score, then most frequent

keywords are listed from the documents. The two example is given in Fig. 4.14 as “post-facto” and “ME”.

Word	Score IfIDF	Frequency IfIDF	Score MRake	Frequency MRake
post-facto	0.00462, 0.00463	2	1, 4, 4, 4	5
Facto	0.036	1	-	-
ঘটনান্তর	0.0204, 0.00327	2	-	-
যন্ত্রকৌশল	-	-	1, 1, 1, 1, 1	5
যন্ত্রকৌশল বিভাগের	-	-	4	1
মেকানিক্যাল	0.00711	1	-	-
ME	0.01452, 0.00732	2	-	-

Fig.4.14: Example of the Frequency of Two Words :Post Facto and Mechanical

#### 4.7 Feature and Pattern Extraction for Decisions with User Query

Then, the process of extracting the query knowledge from these decisions is performed. To extract the desired knowledge, we processed the decision store to get the important text and then with the corresponding features we extracted the knowledge from the text which is mapped with the user query. Fig 4.5. Shows the strategy of the proposed system for keyword extraction and finding user query from the documents. Fig.4.6. presents the algorithm.

In the files, there are many keywords and we have selected the most common and high-frequency keywords from the keyword list as keywords knowledge base for keywords. If a user gives a query from these decision files, at first its look into the keywords knowledge base. The keywords knowledge base has three features specification on it. If the query is not there in knowledge base it will directly be searched for the matching words. The three features where a phrase or keyword is matching with the words in sentences is described below. However, we are not able to find the exact information we are looking for [11], [12] in the direct search. Hence we have used the other features. The features are discussed below.

##### 4.7.1 Content-Based Features

The phrase or user query is directly searched with a set of a regular expression where all the lines which contain the phrase, are fetched by the regular expression (regex). However, in this means we only found out the words exactly matched the phrase and we did not get relevant information regarding our search where only one feature is used. In this case, we have used a set of words called “pratyay” with the query word from the keywords knowledge

base. If the query word is “কমিটি”, so it from the 20 selected keywords. Therefore for every word from selected keywords, there are a set of words with the keyword. Like “কমিটিসমূহ”, “কমিটি-এর”, “কমিটিতে ” etc.

#### 4.7.2 Semantics Features

We found there are some definite words with various kind of synonyms that are used frequently to express the same meaning which are described below.

1. However, in the decision files we found many words in English written in Bangla font. Such as “Sessional” as “সেশনাল” and also as English word. We have tried to make a knowledge base with English word to Bangla font words. So that if the user query is “সেশনাল” then it can find words with “Sessional” also. In the 20 keyword knowledge base we have listed “অর্ডিন্যান্স”, “গ্রাজুয়েট”, “মেকানিক্যাল” , “রেজিস্ট্রেশন” , “কমিটি”, “থিওরী” etc words.

2. Then, there are some English to Bangla words which are used simultaneously. Such as “তত্ত্বীয়” with theory, “কোর্স প্রত্যাহার” with course withdrawal, committee with “কমিটি”, “ঘটনান্তর” with Post Facto, “রেজিস্ট্রেশন” with registration, “মেকানিক্যাল” with Mechanical . They are used as vice versa in all the documents. Therefore we also consider this.

3. Moreover, There are some words with short form and elaboration from. Such as “মেকানিক্যাল” with ME or “এমই”, CASR with “উচ্চ শিক্ষা ও গবেষণা কমিটি” etc. However, these are used simultaneously. We have listed these types of most common keywords for searching.

4. However, a lot of words are there with Bangla synonyms. Such as 'নিম্নলিখিত', 'নিম্নে', 'নিম্নোক্ত', 'নিম্নবর্ণিত', 'নিম্নরূপ' all are same meaning but they are using randomly in the documents. Moreover Bangla words sometimes uses different spelling for same word such as 'ঘটনান্তর' and 'ঘটনান্তোর' etc. For the 20 keywords we have made these knowledge base.

5. We have also classified the same types of words which shows the decision is very significant and they emphasize on the topics. Such as 'জরুরী', 'অধিকতর', 'দ্রুত', 'সতর্ক', 'জরিমানা', 'Compulsory', 'স্বগিত', 'বহুল', 'অঙ্গীকারনামা', 'যতশীঘ্র', 'কঠোরভাবে', 'শীঘ্র', 'গুরুত্বপূর্ণ' and therefore added in the knowledge base. However, so far we have applied many semantic conditions and did not consider the surroundings clues of the sentence. Hence we need Context based features to find out the exact context of the knowledge. Moreover the Bangla

WordNet [57] can be used for Bangla similar words but here we can see from the above discussion that there are many English word written in Bangla font which is pronounced as English. Therefore only Bangla to Bangla meaning and English to English meaning extraction will not give an accurate result.

### 4.7.3 Context-Based Features

In our previous sections, we did not include natural language processing and only by the location of the sentence, we have extracted the knowledge. Here we have considered the context and its connection to the required sentences. Some sentences refer another sentence as a reference or they are connected to a definite topic. From all the decision files in our working domain from the data set, we found five types of connecting words which are given in Fig. 4.15. And discussed below.

1. A connection word list which indicates the specific topics previously described.
2. A connection word list which indicates person/s after a sentence
3. A word list of conditionally connecting words immediately after the sentence used in the dataset.
4. A list of words which immediately talks about future or past with the sentence
5. An explanation list which explains the previous line.

These words are immediately searched in the consecutive next two sentences if the query word is found in the sentence. However, if the query words are found in these connection sentences, then the previous sentence is extracted with the current sentence. These connection words are mostly the first word of the sentence. However, they can be anywhere in the sentence according to the context. So we have considered these words are location independent in the sentence.

### 4.7.4 Classify the Documents with Keywords

All the files with the decision are categorized with the 20 keywords and stored in the files with a meeting date. If the user query is related to these keywords then the information stored in the files is used to extract the knowledge. From the set of total documents, we have extracted the decision making lines. From these lines, we have extracted most frequent 20 keywords in of tf-idf and RAKE. Then with these 20 keywords all feature extraction is done which is described above. Then we have categorized 20 different topics with the 20 keywords. Example of Seven word set with occurrences are given in Fig. 4.16.

Immediate Definite Topics	উল্লিখিত
	উক্ত
	এ ব্যাপারে
	অত্র
	ইহা
	উপরোক্ত
	এই
	এটাকে
Indicating person	তাকে
	তাকে
	তাদের
	তাদের
	তাদেরকে
	তার
Adding extra Condition	তবে
	এছাড়া
	উভয়ক্ষেত্রে
	অন্যথায়
	এছাড়াও
	এতদসাথে
	অপর
Near future or near past	ভবিষ্যতে
	পরবর্তী
	পরবর্তীতে
	অতঃপর
	ইতঃপূর্বে
Explain Immediate line	অর্থাৎ

Fig. 4.15: Connection Word List

Keywords set With Connecting word list	Occurrences
'ঘটনান্তর', 'ঘটনান্তোর', 'Post-facto', 'facto', 'Post Facto'	23
'Post-facto'	7
'Graduate', 'গ্রাজুয়েট', 'স্নাতকোত্তর', 'স্নাতক'	25
'গ্রাজুয়েট'	8
'পরীক্ষা', 'পরীক্ষার', 'পরীক্ষকের', 'পরীক্ষক', 'প্রশ্নপত্র প্রণেতা', 'পরীক্ষক তালিকা'	140
'পরীক্ষক'	18
'ধারা', 'উপধারা', 'দ্রষ্টব্য', 'সংবিধি', 'ধারার', 'নীতিমালটি', 'নীতিমাল'	38
'ধারা'	27
'সারসংক্ষেপ', 'সংশোধিত', 'সংশোধন', 'সংশোধনী'	82
'সংশোধিত'	8
'অর্ডিন্যান্স', 'অর্ডিন্যান্সের', 'Ordinance', 'অর্ডিন্যান্সে', 'অর্ডিন্যান্সটি'	22
'Ordinance'	15
'সেশনাল', 'Sessional', 'থিওরী', 'তত্ত্বীয়', 'theory'	13
'তত্ত্বীয়'	2

Fig. 4.16. Example of Occurrence of Words

## 4.8 Conclusions

In this chapter we elaborate our proposed models within a specific domain of minutes of meeting of academic council of KUET. We effectively demonstrated a system which



automatically discovers and extracts the decisions and agendas from the pool of official Bangla digital documents and allow extraction, detection, and analysis of the documents in the domain. However the finding user query technique can be used in the pure text file. Furthermore, both strategy can be used in any kind of official minutes of meeting with specific features.

## CHAPTER V

### Results and Discussions

#### 5.1 Experimental Setup

In this chapter, we present the experimental results along with the analyses of the proposed schemes. We have analyzed the results of the proposed schemes with precision and recall. Since the process is divided into two parts: “agenda” and “decision” detection, and user query detection, therefore we experimented these two.

All experiments is done on a windows machine having Intel Core i5 2.40GHz processor, 4 GB RAM, and we used Java Python as a programing language to implement our algorithm and Weka [17] as an implementation tool for classification. We have used nltk, matplotlib, scipy, numpy, sklearn, textblob and other python packages. Naive-Bayes Classifier and Gaussian is used to classify the result. For installing pip, we need to download get-pip.py. After that, in the command prompt window and we had to run python get-pip.py

#### 5.2 Performance Analysis of the Structure

The result of the proposed system is discussed in the following sub section.

##### 5.2.1 Extraction of Agenda and Decisions Text Analysis

For result analysis of Extraction of agenda and decisions text, we have experimented with 29 resolutions of Academic Council’s meeting of KUET to extract the data using the described methods in Chapter IV. The Naive Bayes model [10] can manage millions of digital documents efficiently and formerly used effectively by many researchers for extracting sentences and classification [44], [45], [11]. It is defined as below:

If  $F_1, F_2, \dots, F_n$  are the assumed features for sentence  $s \in S$ , Bayes’ rule computes the probability that  $s$  belongs to  $S_d$ , as (3.2), (3.3).

Where,

$S_d$  = set of sentences that are related to the document-element  $d$

$S$  = set of all sentences in the document  $D$ .

We used (3.2), (3.3) to classify the correctly detected sentence by the model with a set of actually desired sentences.

From the 12 Documents, we have collected the content-based features manually. And for all documents, the total number of “*decision*”, “*agenda*” and query sentences are manually counted. Then by the model, we have extracted the desired information with three features. For these three features, 10-fold cross-validation is used. Precision, recall, and F1 are measured to evaluate the result. It is defined as (3.6), (3.7), (3.8) and [12]:

Where,

$T_a$  = Set of data to be detected,  $T_d$  = Set of detected data

$True\ positive = T_a \cap T_d$

$Total\ Predicted\ Positive = T_d$

$Actual\ Positive = T_a$

In the Table 5.1, we have shown the precision, recall, and F1 for random eight documents from the set of 29 document whereas a perfect structure with high precision and high recall mean all results categorized appropriately. Here, we find all the recall are 1 which means all relevant “*decision*” was extracted but how many unrelated “*decision*” was extracted, it cannot to known by this. The precision score tells us most of the “*decision*”

**Table 5.1.** Precision, Recall and F1 for “*Decision*” detection for Random 8 Documents

No	Methods								
	<i>Query-Based : 1</i>			<i>Query and Content-Based : 2</i>			<i>Query, Content and Context-Based : 3</i>		
	<i>Pr</i>	<i>Rc</i>	<i>F1</i>	<i>Pr</i>	<i>Rc</i>	<i>F1</i>	<i>Pr</i>	<i>Rc</i>	<i>F1</i>
1	.879	1	.935	.659	1	.795	.897	1	.945
2	.952	1	.976	.955	1	.977	.857	1	.923
3	.727	1	.842	.786	1	.880	1	1	1
4	.583	1	.737	.917	1	.957	.861	1	.925
5	.806	1	.892	.766	1	.867	1	1	1
6	.893	1	.943	.825	1	.903	.821	1	.902
7	.893	1	.943	.718	1	.836	.964	1	.982
8	.769	1	.870	.929	1	.963	.808	1	.894

extracted is relevant. The precision of eight documents of three methods for mining the “*decision*” with Query-based method (Feature 1), and Query-based with Content-based method (Feature 1 and 2), and Query-based with Content-based and with Context-based method (Feature 1, 2 and 3) is plotted in following Fig.5.1. And here we can find the differences between them where the method is going close to the perfect set of “*decision*” gradually. Here from eight documents we find that, the two documents with three methods (Feature 1, 2 and 3) shows precision, recall, and F1 as 1 which means the accuracy is 100% and one documents with three methods (Feature 1, 2 and 3) is nearly 1. Table 5.2 shows precision, recall, and F1 of a total number of detected “*decision*” for all 29 documents using Feature 1, 2 and 3 and the precision is 87% with 92% F1 measure with the set of exact “*decision*”. And we get the accuracy [12] by the following way for “*decision*”:

$$\text{Accuracy} = \text{All detected Decision} / \text{Total Decision} \quad (5.1)$$

Here, the achieved accuracy is 86% for all 29 documents.

**Table 5.2.** Total Set of Data with Precision, Recall and F1

Total number of Agenda	474			
Total Number of detected Agenda	475			
Total number of Decision	698	Pr	Rc	F1
Total Number of detected Decision	607	.870	1	.920
Total Number of documents	29			

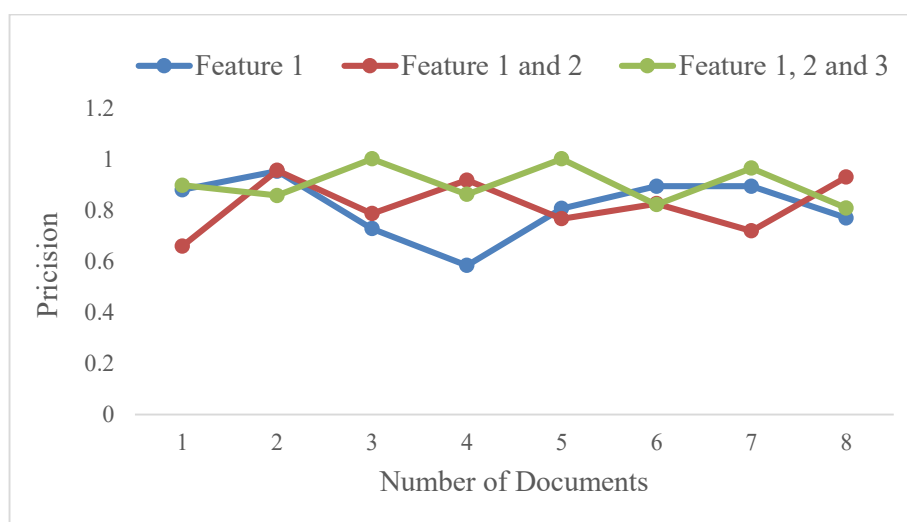


Fig. 5.1: Precisions of the Three Methods of Random 8 Documents

The number of extracting sentences with “decision” using Query-based with Content-based (Feature 1 and 2), and Query-based with Content-based with Context-based (Feature 1, 2 and 3) and the set of a total number of “decision” 29 documents are separately shown in the Fig. 5.2. And we can see how close we are with the perfect set of “decision” for 29 set of data.

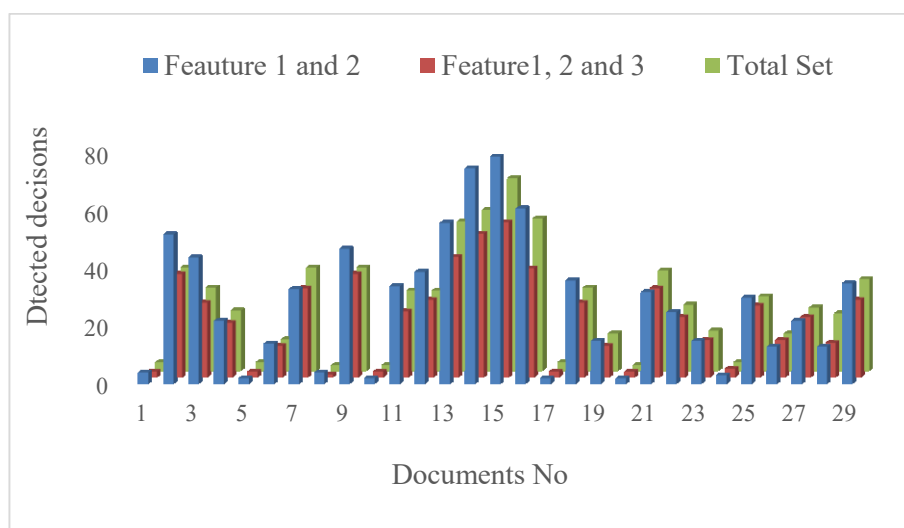


Fig. 5.2: Total Decisions Detected of the Methods- A) Content-Based Method and B) Context-Based Method Merge with the Content-Based Method and C) Total Decision Counted Manually for 29 Documents

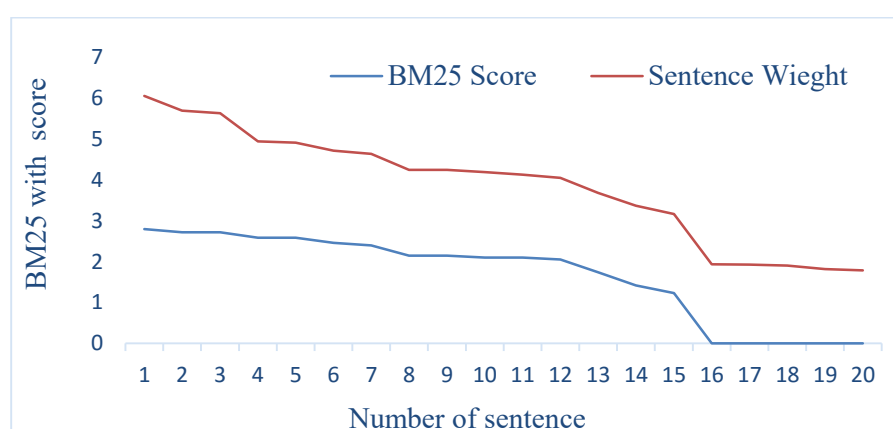


Fig.5.3: BM25 Score and Sentence weight. (The query word- “মেকানিক্যাল”)

Moreover, in 29 documents the query of “Mechanical” as in Bangla “মেকানিক্যাল” is searched and only the one document contains the query with 16 sentences which is shown

in Fig. 5.3. In the Fig. 5.3. it is found that all the sentences except the extracted sentences with the query word, are set to zero in BM25 and the sentence weight is also similar to BM25 after top 16 sentences. Here we have plotted the top 20 sentences, but only 16 sentences have the BM25 score above zero which contains the query.

Moreover, manually we have calculated there are exactly 16 sentences in a single file from a set of 29 documents which actually carries the above keyword. Fig. 5.4. Shows an example of some part of a single document for “decision”.

File =Academic Council-57.pdf  
 Total Number of lines in the document=166  
 Total number of query lines found for “সিভিল”=13  
 Total number of “সিদ্ধান্ত” = 52  
 Total number of “আলোচ্যসূচী” = 22  
 Average length of the sentence in the document=20

খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়  
 গত ১১/১/২০১৮ইং তারিখে অনুষ্ঠিত একাডেমিক কাউন্সিলের ৫৭ তম সভা:

সিদ্ধান্তঃ অত্র বিশ্ববিদ্যালয়ের ৩য় সমাবর্তন উপলক্ষে গৃহীত কার্যক্রম সভায় উপস্থিত সকলে অবহিত হন এবং বিষয়টি ঘটনাত্তোর অনুমোদন করা হলো ।

সিদ্ধান্তঃ অত্র বিশ্ববিদ্যালয়ের ২০১৭-২০১৮ শিক্ষাবর্ষের স্নাতক পর্যায়ে ১ম বর্ষের ওরিয়েন্টেশন ও ক্লাস শুরুর তারিক আগামী ২৫/০১/২০১৮ইং নির্ধারণ করা হলো ।

Fig. 5.4. An Example of “*Decision*” Detection in a Single Document

### 5.2.2 Finding User Query from the Extracted Decision Pool Analysis

For user query analysis, we have experimented with 29 resolutions of Academic Council’s meeting of KUET to extract the data and then using the method collected the decision list. It is divided into three parts: Keywords detection, Query detection, and categorization.

We have also measured the cosine similarity of the sentences and rank them with TextRank algorithm [38], [52] to get the most informative sentences. For the decision sentences, we have to make a similarity matrix to know the top relevance sentences. In the Fig. 5.5. It showed the top TextRank from equation (3.9), (3.10) sentences using cosine similarity where the red words are the keyword extracted by tf-idf and Rake algorithm. Most of these words are from our 20 keyword knowledge base. TextRank also verifies the common frequency words from both tf-idf and Rake. From the TextRank we ranked all the lines and the values are then sent to the Gaussian model from equation (3.11), (3.12), (3.13), and (3.14) to make a classification. We have made three clusters to represent the graph which is shown below in Fig. 5.6.

1. সিদ্ধান্তঃ খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ের পুরকৌশল, তওইকৌশল, যন্ত্রকৌশল, সি.এসই ও ইসিই বিভাগের বিভাগীয় প্রধানগণ কর্তৃক প্রস্তাবিত এবং Co-ordination কমিটি কর্তৃক সুপারিশকৃত একাডেমিক ক্যালেন্ডারসমূহে নির্ধারিত ছুটিগুলো ঠিক রেখে তৈরী করার জন্য পরামর্শ দেয়া হয়।
2. একাডেমিক ক্যালেন্ডারসমূহ সংশোধিত আকারে আবারো অনুমোদনের জন্য আনা হোক।
3. সিদ্ধান্তঃ খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ের পুরকৌশল, তওইকৌশল, সি.এসই ও ইসিই বিভাগের প্রস্তাবিত বিভিন্ন শিক্ষাবর্ষের নিয়মিত পরীক্ষা কমিটি (পরিশিষ্ট-জ্ঞ০৭/০৪ঞ্চ দ্রষ্টব্য) অনুমোদন করা হলো।
4. সিদ্ধান্তঃ খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ের তওইকৌশল ও সি.এসই বিভাগের স্পেশাল ব্যাকলগ পরীক্ষার জন্য গঠিত পরীক্ষা কমিটি অনুমোদন করা হলো (পরিশিষ্ট-জ্ঞ০৭/০৫ঞ্চ দ্রষ্টব্য)।
5. সিদ্ধান্তঃ খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ের ২০০২-২০০৩ শিক্ষাবর্ষের পুরকৌশল, তওইকৌশল, যন্ত্রকৌশল ও সি.এসই বিভাগের বিভিন্ন বর্ষের বি. এস-সি. ইঞ্জিনিয়ারিং পরীক্ষার ফলাফল (পরিশিষ্ট-জ্ঞ০৭/০৬ঞ্চ) ও ডিগ্রী অনুমোদন করা হলো।

Fig. 5.5: Top-ranked Sentences by Textrank

Here we can understand the blue and green groups are overlapping with each other. Therefore there are many lines which holds the same words as keywords hence share the cluster. And the words cosine similarity are also very close to each other as it is a specific domain and most of the topics are expressed with similar words. However, for K-means clustering for decision lines, it made a two-cluster in an unsupervised learning showed in Fig.5.7. It also concludes that these sentences are very similar shown in Fig.5.6.

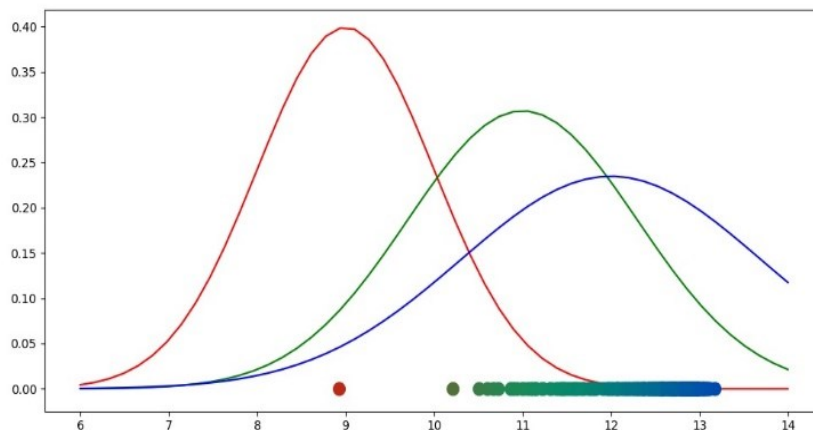


Fig. 5.6: Gaussian Curve for Three Clusters

In the Table 5.3, we have shown the precision, recall, and F1 for decision sentences from the set of 29 document for a single query “মেকানিক্যাল” whereas a perfect structure with high precision and high recall mean all results categorized appropriately. Here, we find the recall are 1. And the precision is 92% with 96% F1 measure. And the accuracy is (5.1).

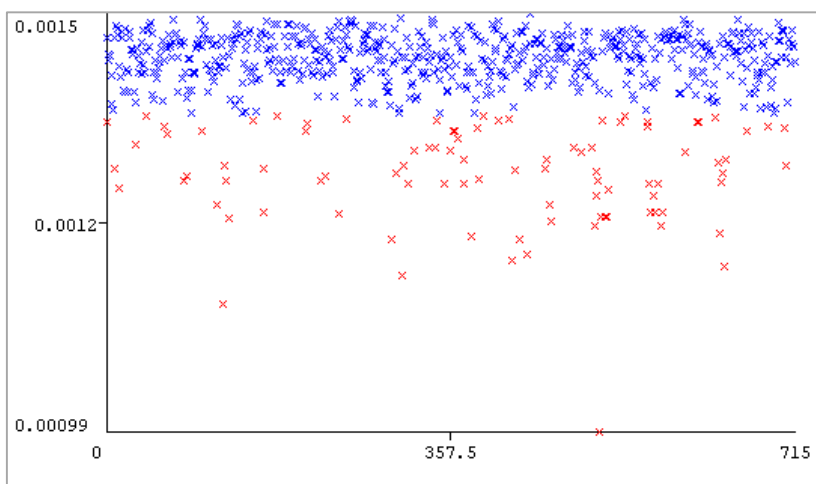


Fig. 5.7: K-Means of the Two Clusters

Then by the model, we have extracted the desired information for different keywords with three features. For these three features, 10-fold cross-validation is used. Precision, recall, and F1 are measured to evaluate the result. It is defined as (3.6), (3.7), (3.8) and [11].

**Table 5.3.** Total Set of Data for One Keyword –মেকানিক্যাল with Precision, Recall and F1

Total Number of ‘মেকানিক্যাল’ related lines from decision lines	93			
Total Number of detected only ‘মেকানিক্যাল’ lines from decision lines [with a single query]	02	Pr	Rc	F1
Total Number of detected ‘মেকানিক্যাল’ related lines from decision lines	87	.925	1	.961

However, in our work from 29 documents the query of “*Mechanical*” as “মেকানিক্যাল” is explored and only in a single document the query is found in 16 lines. But using the described three features here we found all the related words with “মেকানিক্যাল” and only from all the decision pool we got 87 corresponding lines.

### 5.3 Discussion

In this chapter we present the experimental outcomes and analyses of the proposed schemes. As there is no existing method in Bangla knowledge extraction from official documents, we have compared our schemes with the precision and recall and also with the own finding system. We have also made categorization with classification model. In each



case we found relevance result and hence we validated the system. We have showed that the scheme outperforms the exact word extraction schemes. 86% accuracy is achieved for a sample Dataset to detect decisions from the PDF documents.

## CHAPTER VI

### Conclusions

#### 6.1 Summary

Now-a-days large volume of produced Bangla files maintenance and analysis have been a key concern in different aspects of data computing like Big Data. But the margin of large volume is increasing day by day as the required size of data files is expanding gradually. On the other hand, the tools for effective knowledge extraction and finding required data among the large volume is very small as the Bangla language is extremely complex so there are less work done on it. Furthermore, Bangla is the sixth most widely spoken language in the world. So, it is very important to handle large volume Bangla documents efficiently with meaningful way. However, Extraction of the correct information from a huge set of Bangla files is very significant for decision making. The vast amount of legal Bangla files are produced by professionals, Judicial, and academician and is very difficult to find the previous or specific information.

In this research we have described a new model, a domain specific information extraction system for Bangla official documents where the user will find his/her desired information using a set of the various knowledge-based method and find the decisions and the topic of discussions which are taken the meeting written in Bangla text. In this work, we have presented the semantic and other features with natural languages processing. The results are presented here with a precision and recall which showed that the proposed algorithm achieved a high performance. The accuracy for this sample Dataset for “decision” extraction is 86%. The knowledge is also classified with keywords from the documents. However, the major disadvantage of the system is that knowledge base is for a specific domain with a small dataset and it do not extract information from the tabular data. Moreover, the keywords knowledge

base is small and do not extract information using vocabulary and morphological investigation of words.

## **6.2 Recommendations for Future Works**

Since the proposed model is applied in the official Bangla PDF AC files from KUET for knowledge extraction, text, html or doc can use the scheme. More specifically –

- This scheme can be applied to Big Data.
- One important future direction of the work is that; the scheme can be easily implemented for discovering various patterns with semantics analysis.
- It will be very efficient to apply synopsis generation, summarization for a specific domain.

## REFERENCES

1. Weiss, G.M. and Davison, B.D., Data Mining. Handbook of Technology Management, H. Bidgoli, 2010.
2. PDF in 2016: Broader, deeper, richer, <https://www.pdfa.org/pdf-in-2016-broader-deeper-richer>, Accessed on August 10, 2018
3. Staar, P.W., Dolfi, M., Auer, C. and Bekas, C., Corpus Conversion Service: A machine learning platform to ingest documents at scale. arXiv preprint arXiv:1806.02284, 2018.
4. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., From data mining to knowledge discovery in databases. AI magazine, 17(3), p.37, 1996.
5. Baker, J.B., Sexton, A.P., Sorge, V. and Suzuki, M., September. Comparing approaches to mathematical document analysis from PDF. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 463-467). IEEE, 2011.
6. Zanibbi, R. and Blostein, D., 2012. Recognition and retrieval of mathematical expressions. International Journal on Document Analysis and Recognition (IJDAR), 15(4), pp.331-357, 2012.
7. Mandal, S., Chowdhury, S.P., Das, A.K. and Chanda, B., Automated detection and segmentation of table of contents page from document images. In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on (pp. 398-402). IEEE, 2003.
8. Wu, Z., Das, S., Li, Z., Mitra, P. and Giles, C.L., Searching online book documents and analyzing book citations. In Proceedings of the 2013 ACM symposium on Document engineering (pp. 81-90). ACM, 2013.
9. Chiu, P., Chen, F. and Denoue, L., Picture detection in document page images. In Proceedings of the 10th ACM symposium on document engineering (pp. 211-214). ACM, 2010.
10. Kataria, S., Browner, W., Mitra, P. and Giles, C.L., Automatic Extraction of Data Points and Text Blocks from 2-Dimensional Plots in Digital Documents. In AAAI (Vol. 8, pp. 1169-1174), 2008.

11. Bhatia, S. and Mitra, P., Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems (TOIS)*, 30(1), p.3, 2012.
12. Tuarob, S., Bhatia, S., Mitra, P. and Giles, C.L., AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data. *IEEE Transactions on Big Data*, 2(1), pp.3-17, 2016.
13. Das, A. and Bandyopadhyay, S., Theme detection an exploration of opinion subjectivity. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1-6). IEEE, 2009.
14. Islam, M.S., Research on Bangla language processing in Bangladesh: progress and challenges. In *8th International Language & Development Conference* (pp. 23-25), 2009.
15. Molla, M.K.I. and Talukder, K.H., Bangla number extraction and recognition from document image. *5th ICCIT 2002*, pp.200-206, 2002.
16. Sarkar, K., Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240, 2012.
17. Weka, <https://www.cs.waikato.ac.nz/ml/weka/>, Accessed on August 10, 2018
18. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S. and Kanahori, T., INFTY: an integrated OCR system for mathematical documents. In *Proceedings of the 2003 ACM symposium on Document engineering* (pp. 95-104). ACM, 2003.
19. Baker, J.B., Sexton, A.P. and Sorge, V., A linear grammar approach to mathematical formula recognition from PDF. In *International Conference on Intelligent Computer Mathematics* (pp. 201-216). Springer, 2009.
20. Baker, J.B., Sexton, A.P. and Sorge, V., Faithful mathematical formula recognition from PDF documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (pp. 485-492). ACM, 2010.
21. Anderson, R.H., Syntax-directed recognition of hand-printed two-dimensional mathematics. In *Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium* (pp. 436-459). ACM, 1967.
22. Mandal, S., Chowdhury, S.P., Das, A.K. and Chanda, B., Automated detection and segmentation of table of contents page from document images. In *Document Analysis*

- and Recognition, 2003. Proceedings. Seventh International Conference on (pp. 398-402). IEEE, 2003.
23. Pdfbox, <http://pdfbox.apache.org/>, Accessed on August 10, 2018
  24. <https://www.snowtide.com/>, Accessed on August 10, 2018
  25. <http://www.xpdfreader.com/about.html>, Accessed on August 10, 2018
  26. <https://www.pdflib.com/products/tet/>, Accessed on August 10, 2018
  27. Bhatia, S., Mitra, P. and Giles, C.L., Finding algorithms in scientific articles. In Proceedings of the 19th international conference on World wide web (pp. 1061-1062). ACM, 2010.
  28. Das, A. and Bandyopadhyay, S., Phrase-level polarity identification for Bangla. *Int. J. Comput. Linguist. Appl.(IJCLA)*, 1(1-2), pp.169-182, 2010.
  29. Das, A. and Bandyopadhyay, S., Sentiwordnet for bangla. Knowledge Sharing Event-4: Task, 2, pp.1-8, 2010.
  30. Bhattacharya, U., Parui, S.K. and Mondal, S., Devanagari and bangla text extraction from natural scene images. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 171-175). IEEE, 2009.
  31. Hassan, A., Amin, M.R., Mohammed, N. and Azad, A.K.A., Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models. arXiv preprint arXiv:1610.00369, 2016.
  32. Ramanathan, A. and Rao, D.D., A lightweight stemmer for Hindi. In the Proceedings of EACL, 2003.
  33. Rose, S., Engel, D., Cramer, N. and Cowley, W., Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, pp.1-20, 2010.
  34. Engel, D.W., Whitney, P.D., Calapristi, A.J. and Brockman, F.J., Mining for emerging technologies within text streams and documents (No. PNNL-SA-64618). Pacific Northwest National Lab.(PNNL), 2009.
  35. Whitney, P., Engel, D. and Cramer, N., Mining for surprise events within text streams. In Proceedings of the 2009 SIAM International Conference on Data Mining (pp. 617-627). Society for Industrial and Applied Mathematics, 2009.
  36. Mihalcea, R. and Tarau, P., TextRank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, 2004.
  37. [https://github.com/vgrabovets/multi\\_rake](https://github.com/vgrabovets/multi_rake) , Accessed on Nov 10, 2018

38. Pay, T., Lucci, S. and Cox, J.L., An Ensemble of Automatic Keyphrase Extractors: TextRank, RAKE and TAKE.
39. Lynn, H.M., Lee, E., Choi, C. and Kim, P., Swifrank: an unsupervised statistical approach of keyword and salient sentence extraction for individual documents. *Procedia Computer Science*, 113, pp.472-477, 2017.
40. Cleveland, H., Information as a resource. *Futurist*, 16(6), pp.34-39, 1982.
41. Porter, M.F., An algorithm for suffix stripping. *Program*, 14(3), pp.130-137, 1980.
42. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M., Okapi at TREC-3. *Nist Special Publication Sp*, 109, p.109, 1995.
43. Schütze, H., Manning, C.D. and Raghavan, P., Introduction to information retrieval (Vol. 39). Cambridge University Press, 2008.
44. Kupiec, J., Pedersen, J. and Chen, F., July. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 68-73). ACM, 1995.
45. Teufel, S., Sentence extraction as a classification task. *Intelligent Scalable Text Summarization*, 1997.
46. <https://medium.com/@starang/precision-and-recall-a-brief-intro-38589a21a09>, Accessed on August 10, 2018
47. [https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix), Accessed on August 10, 2018
48. <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9> Accessed on August 10, 2018
49. <https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>, Accessed on August 10, 2018
50. Skabar, A. and Abdalgader, K., Clustering sentence-level text using a novel fuzzy relational clustering algorithm. *IEEE transactions on knowledge and data engineering*, 25(1), pp.62-75, 2013.
51. Mihalcea, R., Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (p. 20). Association for Computational Linguistics, 2004.
52. Brin, S. and Page, L., The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7), pp.107-117, 1998.
53. Erkan, G. and Radev, D.R., Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22, pp.457-479, 2004.

54. Gope, M. and Hasehm, M.M.A., Knowledge Extraction from Bangla Documents: A Case Study. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE, 2018.
55. Hassan, T., September. Object-level document analysis of PDF files. In Proceedings of the 9th ACM symposium on Document engineering (pp. 47-55). ACM, 2009.
56. Shams, R., Hashem, M.M.A., Hossain, A., Akter, S.R. and Gope, M., Corpus-based web document summarization using statistical and linguistic approach. In Computer and Communication Engineering (ICCCE), 2010 International Conference on (pp. 1-6). IEEE, 2010.
57. <http://indradhanush.unigoa.ac.in/public/webcontent/webcontent.php?id=37>, Accessed on Nov 10, 2018