Automatic Sinhala Text Summarization for Government Gazettes using Abstractive and Extractive Methods

H. M. R. Y. Jayawardane

2021

Automatic Sinhala Text Summarization for Government Gazettes using Abstractive and Extractive Methods

A dissertation submitted for the Degree of Master of Computer Science

H.M.R.Y. Jayawardane
University of Colombo School of Computing
2021

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: H. M. R. Y. Jayawardane
Registration Number: 2017/MCS/040
Index Number: 17440402

_____ _____

Signature of the Student Date

This is to certify that this thesis is based on the work of Mr. H. M. R. Y. Jayawardane under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,
Supervisor Name: Mr. W. V. Welgama

_____ _____

Signature of the Supervisor Date

ABSTRACT

During this new era, information is very accessible and the amount of text data from various sources has increased dramatically. However, most of them can distract the reader from the most important information due to using larger paragraphs, examples, complex arguments, grammar, and some vocabularies. Since time is one of the most important facts in the 21st century, people want to summarize these contexts and retrieve only the important information in a shorter time.

The Gazettes are important to people in different way and there was no attempt on summarization solution for the area, this research emphases on the summarizing gazettes in the Sinhala language. This research solution is to provide summarized output for Sinhala gazettes by identifying the most important and relevant sentences based on linguistic and statistical features of a given text, using an abstractive and extractive approaches. Even though there are very few attempts done on the Sinhala Summarization this is the first attempt on summarizing Sinhala gazettes.

The project was evaluated by machine summaries with the summaries created by the author. The system has been tested with 450 actual Sinhala gazettes and final results were attached in the Appendix section. Further, this provides a turning point for future researches on automatic text summarization in Sinhala language.

ACKNOWLEDGEMENT

I am especially grateful to Mr. W. Viraj Welgama, senior lecturer of the University of Colombo School of Computing, who have been cooperative for this project and who worked energetically to provide me ideas information to pursue project to a possible solution. This work would not have been possible without his support.

No one was more important to me on this path than my family members. I would like to thank my parents, whose love and advice accompany me in everything I am looking for. These are the best models that provide endless inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NLP	Natural language processing
PDF	Portable Document Format
UI	User Interface

CHAPTER 1
INTRODUCTION

The purpose of this chapter is to give an overview of Sinhalese automated text summarization for official Gazettes using abstract and extracted methods. It begins with a problem domain, brief introduction to the abstractive and extractive summarization, Sinhala text summarization, Sri Lanka Gazettes structure and the existing boundaries challenged by Sinhala Text Summarization. In addition, this section presents the purpose, aims, scope and characteristics of the projected solution.

1.1 Motivation

Much research has been carried out over the last six decades to adopt technology to most common languages such as English, because the technology was born with these languages. Later, when the other languages were enabled with the technology, the first step was to apply the existing techniques and findings to these languages rather than reinventing techniques for the same issues. This helps such languages to adapt to the technology rapidly in shorter time and less cost while it also helps linguists to identify the language families based on the adaptability. This scenario also motivated the author to apply such existing summarizing techniques to Sinhala and find their applicability to languages such as Sinhala. In Sri Lanka, the documents that must be prepared in the three official languages: Sinhala, Tamil and English. For example, government documents, gazettes, public notices, etc. they are published in all three languages. This is an area where machine translation can be of great use, especially when it concerns a specific area.

1.2 Statement of the problem

A summary is a text that consists of sentences that contain important data in the original transcript in a briefer form. The biggest benefit of using a summary is that it decreases analysis time. Text summarization methods can be divided into extractive and abstractive summarization. The extractive summarization is to select important sentences and paragraphs from the original document and then combine them into a briefer form. Abstractive summarization is the understanding of basic concepts in a document and then expressed in an understandable natural

language. The text summarization is divided into two main groups: indicative and informative [1]. Inductive summarization only shows the user the main idea with the text. The typical size of this summarization is five to ten percent of the original document. The informative summarization provides accurate information about the text. The output makes less than 3o% of the original context.

During this new era, information is very accessible and the amount of text data from various sources has increased dramatically. However, most of them can distract the reader from the most important information due to using larger paragraphs, examples, complex arguments, grammar, and some vocabularies. Since time is one of the most important facts in the 21st century, people want to summarize these contexts and retrieve only the important information in a shorter time.
In this context, there is an urgent need to develop an improved mechanism for extracting the most important data and information from documents efficiently and accurately. With today's rapid growth of using smartphones, tablets, and electronic devices this has become a requirement for the development of automated tools to summarize the content of the documents and show them to users [2].

Automatic text summarization creates brief and comprehensible summaries without human help while protective the meaning of the original document. It is very difficult because when people summarize a text, we usually read the whole text to improve considerate and then publish a summary by importance its main facts. Because machines are lack of language and human skills, automatic text synthesis has become a actual problematic and extraordinary task.

Although automatic text summarization is not a new area of research, it has received significant attention in the research community in recent years. Automatic text synthesis has become an important and useful area of research in natural language processing and information retrieval. Most modern methods of text synthesis are very useful for understanding context. These mechanisms can be triggered by reading human word memory, relationships between words, and cognitive processes. In the reading process, a change in human memory of words is used to indicate the meaning of the context according to the sentence, and then the sentence is ordered and extracted to form a summary.

We realize the importance of a summarization and how it can help you remember things. In this project, we enforced to use this approach for Sinhala Gazettes. Sinhala is one of the indigenous and national language in Sri Lanka. Even though Sinhala is the main language in Sri Lanka and spoken by 19 million people, yet, there are few studies on it in computational linguistics about this area [3]. Out of those, there is a very few attempts done on the Sinhala Summarization. In general, since the resources of the Sinhalese are limited, great efforts are needed. When the required level of precision is reached, it will be a milestone in the natural language processing of the Sinhalese language.

A government gazette also known as official gazette is a periodical publication that has been authorized to publish public or legal notices. It is usually established by statute or official action and publication of notices within it, whether by the government or a private party, is usually considered sufficient to comply with legal requirements for public notice [4].

1.3 Research Aims and Objectives

These Gazettes documents are containing one or several pages and usually the context is having in large sentences. Due to that reason many of Sri Lankans are may have difficulties to find and understanding the correct information quickly.

Since Gazettes are important to people in different way and there was no attempt on summarization solution for the area, this study attentions on the problematic of summarizing Gazettes in the Sinhalese language. This is one of the studies to summarize the text in the Sri Lankan language. In short, a lot of effort must be made due to the reduced availability of resources for the Sri Lankan language. If the desired level of precision is achieved, it will be an important step in NLP in the Sri Lankan language. The foremost objective of this study is to find the most suitable method to summarize the Sinhala text. One of the objectives of the project is to apply suitable methods, technologies, and construction to determine its adaptability. The tricky part of the project was finding a way to improve accuracy and performance without changing the sense of context.

1.3.1 Project Aim

The aim of the project is to summarize a long context in Sinhala Gazettes by identifying the most important information and output it as more organized approach.

1.3.2 Project Objectives

The objective of the project is to use the concepts of automatic text summarization and create a prototype to output the summarization document for Sinhala Gazettes. Including that, the following will be the project objectives that this project is going to fulfill.

- **Convert and summarize complex Gazettes into quick understandable document**

  As mentioned in the "Research Problem" area, the government gazettes carry out notices which are issued by the government and important to a lot of people. The first and main objective of the project is to build a prototype to summarize those gazettes and show only the important contexts without going through the whole document. There were no previous attempts recorded for developing an automated summarizer for any kind of gazettes or Sri Lankan gazettes. The proposed approach will use both abstractive and extractive summarization methods to fulfill the requirement. More details will be covered in the "project scope" section.

- **Find and apply techniques that are suitable and adaptable to the Sinhala language summarization.**

  This project main object is to apply suitable procedures, approaches, and construction to find the adaptableness of them. The interesting part of this project is to find a method that doesn't need many philological possessions to accomplish accurateness and performance.

- **Provide a benchmark for automatic Sinhala text summarization for future researches on the Sinhala language.**

  There are a large number of NLP, text summarization, wordnet, and other related research carried out for rich languages such as English. However, Sinhala is considered a

less-resourced language in the field of NLP and there were a handful of previous attempts available in the Sinhala text summarization. This project will help future researchers find better ways to do the Sinhala summarization.

1.4 Scope

This project work is a wider scope to build a summarization using automatic text summarization and combine with the abstractive-based and extractive-based approaches for the source text which is a lengthy gazette notice. This project deals with the automatic summarization abstract and extractive methods, and techniques which recommend to use in summarization. Automatic text summarization needs to be proven effective, accurate and able understand the core context. To improve the efficiency and effectiveness of the application, process needs to focus only on the summary that includes the appropriate information provided and improved context of the information being processed. A summarization system should be designed taking correct approach what excites a particular rule and how they perceive information.

When the user uploads the gazettes file as a PDF format, the tool deals with the automatic summarization by using the abstract methods, extract methods, techniques, and will generate a summarized document as the output. The proposed project will use the template base method as the abstract approach to filter out the important details into the selected sections. Also, use the extractive methods to summarize the lengthy paragraphs.

1.5 Structure of the Thesis

In this chapter it explained the brief introduction to the project, the problem domain, problem definition of the application with objectives, scope and project features. A basic idea of what we are doing and what we are trying to accomplish with this were address though this chapter.

CHAPTER 2
LITERATURE REVIEW

This section will observe on the present literature existing on producing text summarization from abstractive and extractive text summarization by NLP approaches for Sinhala Gazettes. The purpose of the literature review is a detailed study conducted to identify current state-of-the-art technology, identify and evaluate existing technology, approaches used in former studies, and determine the most suitable method. Also, the restriction, pros and cons of the current methods have been deliberated in this chapter.

2.1 Automatic Text summarization Overview

In today's world, creating accurate and intelligent summaries of long documents and texts is a popular study and a challenge for the industry due to time constraints. This number has increased in recent years due to the growing of the Internet, and people are overcome by the number of documents and information on the Internet [7].

According to a study [9], summaries are defined as "text that contains important information from the source, does not exceed half of the original text". By predominant the key perceptions and the information, constructing a good and exact summary is called automatic text summarization [7]. Automatic text summarization has been advanced and used in many areas in past years. For example, search engines are specific. Use website summary techniques to create headlines, such as news-based content.

Automatic text composition is very difficult because when a person is summarizing a text, the usual procedure is to read all the text or document and write important concepts to deepen understanding. The quality of human-drawn summary may be good, but it takes a long time [10]. Integration this is a difficult because computers are lack of human skills such in thinking, creativity and linguistics [7].

In the automatic text summarization, there are three important characteristics of study which are defined by the explanation [10].

1. Summaries which are produced from a single document or multiple documents
2. Summaries should preserve important information
3. Summaries should be short

There are basically two main methods to automatic text summarization. Abstract and extractive. The extractive method extracts sentences and words from the source text to create a summary. The abstract method, examines linguistic expressions and uses linguistic techniques to generate more human-generated summaries [8].

2.2 History of Automatic Summarization

Computer text summarizing experiments began in the late 1950s by explaining a surface-level approach. This work can be considered the first mathematical work on automatic extraction. The use of subject characteristics such as word frequency has made a good start in the study of automatic summarization.

In the early 1960s, researchers began using a basic analytical approach. The use of sentence position feature was introduced in this area by Edmundson in 1969, who used three other features: keywords, heading words and the sentence location, in addition to word frequency. He found that the combination of display phase, headword, and sentence position was the best feature for text summarization. He also said that location is the best personality trait and only keywords are the worst. In the early 1970s, interest in this area renewed and the first commercial application for automatic abstracting was developed. Pollock and Zamora have developed an automated compiler for the Chemical Abstracts Service (CAS). This compiler used key phrases that were primarily specific to the chemical subdomain and then used it as a commercial product [12].

In the late 1970s, a broader entry-level approach was used. First, they experimented with a speech approach based on the narrative grammar of the time. Entry-level AI-based approaches such as scripts, logic and production rules, semantic networks, and some hybrid approaches used in the 1980s [12].

By the late 1990s, the field of automated synthesis had been actively developed and all types of approaches had already been considered due to commercial and government interest in

applications. The study is now focused on excerpts rather than summaries, renewing interest in previous approaches at the superficial level. However, there is a movement to focus on automatic summarization with a focus on creating natural language, and in this area, instead of focusing on individual data, multi-document summarization, multilingual summarization, we are exploring new areas such as multimedia summarization.

2.3 Abstractive Text Summarization

Creating a short summary of some sentences or headings that reflect the foremost awareness of the text of the article is called abstract text. Abstraction is frequently done by mapping the sequence of contribution words in the source document to the sequence of target words [9]. Associated to extractive summarization, the abstract summarization is an efficient way to generate a precise summary of information since the summary extracts data from numerous possibilities [8]. Abstract summarize information in an easy-to-read and grammatically precise format. Abstract composition can be alienated into two core fragments,

- Structured approach
- Semantic approach

2.3.1 Structure Based Approach

This structure can always be categorized using psychological schemas such as models, ontology, rules, and alternative structures such as tree and graph structure.

2.3.1.1 Template based method

The template-based method serves as a guide for viewing documents or text. Here we use language model extraction rules to specify the summarization. It first, group human-authored summaries and apply multiple alignments to them to create a template. The meeting transcript then identifies a group of human statements that explain certain aspects and emphasize relationships that are worth summarizing. It identifies information related to the subject of the input document and converts it into a database.

2.3.1.2 Tree based method

Dependent on the details of the article or transcript, this method generates a dependency tree. By locally aligning multiple sequences from bottom to top, this method had better be able to recognize a general set of information. Entering numerous articles or text to procedure these entries determine the fundamental theme. Sentence classification is complete when the topic is completed using the text alliance algorithm [23].

2.3.1.3 Ontology based method

Each domain represents its particular ontology and has its individual data construction. Furthermost of the papers available on the Internet are field related. One of the main models used to illustrate this method is "fuzzy". In fuzzy ontology, the fuzzy inference phase produces a degree of membership. Different ontological events result from their vague ideas.

2.3.1.4 Rule based method

By selecting the most efficient and important parts of the produced information, this data abstraction regulation corresponds to single or more aspects of the group. One method is to find semantically relevant nouns and verbs in the information provided. Another approach is to summarize abstract text using word graphics, phonetic rules, and syntactic restrictions. The statement reduction phase is founded on various aspects such as keywords, syntactic constraints, and input statements. Another approach to abstract text synthesis is text synthesis based on random forest classification and merit evaluation. In this method, a cross-validation classifier is trained to calculate a characteristic score for preprocessing the data. This classifier determines if a proposal belongs to a resume. All selected offers are generated based on minimum redundancy and maximum relevance.

2.3.2 Semantic Based approach

This allows you to characterize the content of a text resource according to the semantic domain rather than the usual set of words. The ultimate goal is to take advantage of many sensory differences. This allows you to identify the information that is actually displayed in the context.

2.3.2.1 Multimodal semantic method

This approach uses selected metrics to capture and score all key ideas and relationships between them. Then rank the selected ideas and generate the summary.

2.3.2.2 Information item-based method

This method produces a summary from an abstract representation of the delivered document instead of suggestions from the exact content. It is important to more generalize the identification of all text units, attributes, predicates between them, and predicate characteristics.

2.3.2.3 Semantic graph method

For source code, this approach uses the creation of a semantic graph called Rich Semantic Graph (RSG) to perform the composition. The final summary is generated from the reduced semantic diagram.

2.3.2.4 Semantic text representation model

Instead of parsing the syntax and structure of the text, this approach ranks the most important predicate arguments and selects the context. The final summary is created using a language generator.

In the absence of a generalized structure, summarizing abstract text can be a serious problem because it is difficult to analyze and agree on an analysis tree [12]. Extracting important information and executing sentence order is always an open topic in text abstraction. Moreover, it is difficult to abstract and summarize the compression by paraphrasing and reformulation, and vocabulary substitution. Extracting summaries usually gives better results than abstractive summarization [9]. In fact, in data-driven methods like sentence mining, abstract text summarization technology will face problems like inference, semantic representation, and language generation, which are relatively more difficult. [23].

2.4 Extractive Text Summarization

The basic approach to summarizing snippets is to select the furthermost weighty words or phrases from a particular text before creating the summarization [14]. This approach first gives

high scores to individual words in the source text, then evaluates high-scoring sentences and words. All words and phrases are copied directly from the source. This method cannot generate new words that are not in the original text. Extraction resumes can be broadly divided into three categories.

- Represent the main aspects of the text, create an in-between depiction of the original text.
- Evaluate your suggestions according to your presentation.
- Select the proposal summary with the highest rating

2.4.1 Intermediate Representation

Each summarizer uses an intermediate representation to try to find the content presented in the original text, primarily based on that depiction. Subject presentations and metric presentations are the foremost methods used in this presentation. The subject view uses a thematic approach to words, Bayesian subject models, latent semantic analysis, and more to interpret the text described in the source text. In the indicator view, each sentence was marked as a list of characteristics based on a particular sentence, sentence length, etc. [23].

2.4.2 Sentence Score

Subsequently creating the in-between depiction, evaluate each sentence constructed on the importance of the original text. In topic presentations, a topic created by adding evidence of various indicators creates a score by calculating how well the main topic is explained [23].

2.5 Text Rank Algorithms

Graphical analysis such as the Google PageRank algorithm and Kleinberg's algorithm has proven to be very successful in the areas of dating, social networking, and analysis of link structure on the World Wide Web. This procedure can be used to create an extractable resume, extract keywords from a specified paragraph, and clarify terminology.

The elementary impression behind the TextRank model is "elective" and "recommending". If a apex is connected to additional apex, it is distinct as "matching". The more votes a particular apex gets, the more important it becomes [23].

This model, the figure is created from NL text and contains partial or numerous connections between apexes removed from the text. The TextRank procedure consists of the following basic stages:

1. Use tokenization to identify blocks of text and add them as apexes in the diagram.
2. Determine the relationships between the apexes in the diagram and draw the edges between the apexes.
3. Repeat the graphics-based text classification algorithm until it converges.
4. Place corner points according to the final score.

2.5.1 Keyword extraction using TextRank

Identifying the superlative keywords that define the contented of an article is named as keyword extraction. This keyword extraction technique splits the first specified transcript into words. Formerly the stop words need to be found and proceeded. Altogether vocabulary units are then conceded to determine the currency of a particular word [23].

Lastly, all the shortened text is supplementary to the objectiveless and graphs. Respectively apex in the figure embodies a few vocabulary units from the original source [15]. The relationship between dual vocabulary substances is termed a link [23]. Afterward the diagram is created, each apex starts with an original value of 1.0. The procedure runs in about 30 repetitions until it congregates to a specific edge of 0.0001.

The vertices are sorted after getting the final estimate for each node. The most important keywords are selected based on the length of the original document. If the document is long, more keywords will be selected, and if the original document is short, fewer keywords will be selected.

2.5.2 Sentence Extraction using TextRank

Sentence extraction procedure is comparable to the keyword extraction procedure in that both methods try to determine the furthermost representative keyword phrase for a particular text.

This technique adds the highlighted vertices to the weighted graph for each code sentence. The score is calculated based on the similarity between the two sets [15].

$$Similarity\ (s_i s_j) = \frac{l\ \{\omega_k \mid \omega_k \in s_i\ \&\ \omega_k \in\ s_j\ \}\ l}{\log(|s_i|) + \log(|s_j|)}$$

Where S_i and S_j are the given two sentences and words that appear in the sentence are $S_i = w1_i$, $w2_i$, $w3_i$... wn_i. In the weighted graph, the assigned score for each node indicates the strength of the connections established between various sentences in the source text. After the algorithm is executed, the top sentences are selected based on the final generated score for each one [23].

2.5.3 TextRank usage in other languages

As stated by the Liu [10] they have used an enhanced TextRank algorithm to do the text summarization in Tibetan language.

$$S\ (V_i) = (1 - d) + d * \sum_{j \in \ln\ (V_i)} \frac{1}{|out\ (V_j)|}\ S(V_j)$$

Where the S(v) represents the weight of each node, B(v) is a node set which point to node V. "d" is the damping factor which vary between 0 to 1 and F(u) is node set which is pointed by node v. [10] The following figure displays the basic theory of "vote" and "recommend". When there is a link between point A and point B, then the point A is recommended to point B. If the point B gets more votes it is considered that the point B is more important [23].

To extract keywords from specific Tibetan text, they used the TF-IDF algorithm. In this method, each word is assigned a weight and the word with the highest rank is selected based on the weight [23]. To create an automatic resume in Tibetan, follow these steps:

1. Compare the importance of the ranked keywords and highlight the three main keywords
2. Identify the sentence that contains the keyword selected above. If a phrase containing the three keywords selected above is found, that phrase is selected as a summary.

If not, look for a phrase that contains the two most important keywords. Otherwise, look for a sentence that contains the first number one keyword.

2.6 WordNet

A great vocabulary dataset designed for a particular linguistic is called WordNet. All nouns, adjectives, adverbs, and verb synonyms are collected on WordNet. Synonyms are linked together in WordNet by important theoretical semantic and philological relationships. Extracting common Sinhala from the corpus and seeking expert advice creates the Sinhala WordNet for a fusion approach [16].

Recently, a new Sinhala language, WordNet, was introduced. This Sinhala WordNet is based on the English WordNet (Princeton). This approach uses the Hindi WordNet for research [17]. The development of a fully functional and comprehensive Sinhala WordNet is seen as an important step in approaching Sinhala NLP, such as information retrieval systems, Sinhala text synthesis systems, Sinhala text classifiers, and Sinhala translators. can do. This WordNet has noticed that the words are thin, but there are important differences in the format of the written and spoken words. In a nutshell, Sinhala does not specify any form of gender, which is usually male and female. The gender of the noun is used to determine the form of the verb [17] [23].

Sinhala coined words can be divided into three categories. They are native words, words in another language without change (තත්සම-thathsama), and words in another language with change (තත්භව-thathbawa). The inherited words come mainly from English, Pali, Hindi, Tamil and Portuguese.

When writing sentences in Sinhala, the source of the word must be taken into account. The Sinhala verb called vibhakti (විභක්ති) has nine morphological forms. In addition, compound words called "Sandi" (සංධි) and "Samasa" (සමාස) are formed. The etymology of words is the basis for creating these shapes. Therefore, when you save a word to WordNet, the most common morphological form is saved with the root word. As a resource to support language processing tasks, building Sinhala is actual significant and necessitates important expertise and resource sharing [16].

2.7 Background Study and Similar researches

The foremost impartial of this study is to find the furthermost appropriate method to summarize the Sinhala text. Here limited amount of earlier efforts are existing in the Sinhala text summering. Those studies have been different constraints. Anyhow Considering the NLP and automatic text

summarization there are a fairly quantity of researches have been done for rich languages such as English, Spanish and Hindi. Different approaches that have been taken in these researches to summarize text and how and what are the applicability of such approaches for less resourced languages such as Sinhala.

2.7.1 Automated Text Summarization for Sinhala

This study focuses on some of the classic methods that attempt to use the most appropriate method to identify key information in Sinhalese for an accurate summary. To benefit from all these characteristics, this study suggests the best possible linear combination of the identified characteristics [11].

The proposed method is evaluated by comparing machine-generated and hand-drawn summaries, and primary assumption is based on those man-made summaries are perfect. The results show that the phrase search function is the best single function to extract the most informative phrases from Sinhalese articles, while the linear combination of keywords, title and phrase search functions is the best. The result shows some equations that govern the flow of information in Sinhalese and can be used in many similar applications.

2.7.2 Automatic summarization of scientific articles

The study thoroughly tested an advanced system of summarizing scientific articles. Covers all aspects of complex tasks, including solutions, scoring, and the corpus used in the scoring process. They also highlighted some of the pros and cons of this method. Their research showed that the combination of extraction methods, single article summarizing, statistical methods (TF-IDF) and machine learning (SVM, Naive Bayesian analysis and clustering) and analysis methods have strong advantages [13].

The main associated problems include the unavailability of a basic corpus for teaching and testing, a basic summary for comparison, appropriate assessment indicators, and a necessary basic framework for comparison. Although graphical-based methods have successfully solved the problem of multi-document summarizing, they had attention in the field of automatic summarizing of scientific articles. From single article summarizing, from multi-article retrieval to abstract, more research is needed to expand knowledge in this area to improve consistency and readability of the

result. Deep learning techniques are also worth learning because researched explained that they can successfully solve other complex NLP problems.

2.7.3 Neural Extractive Text Summarization with Syntactic Compression

In this work, they proposed a neural network framework that uses a rules-based format to extract and summarization compression. Their model consists of a sentence extraction model attached to a compression classifier that decides whether to remove a grammar-based compression parameter for each sentence. The goal of training a model is to find a set of predictable results for the extraction and compression solutions, and by combining heuristics and ray search methods. Model is superior to previous work in the Journal Corpus, has made significant progress on the extraction model, and appears to have acceptable grammar based on human judgment. Also, the performance of this method is better than that of the standard compression engine. Manual evaluations indicate that the output of the model is often remains grammatical [18].

2.7.4 Automatic text summarization of legal cases: A hybrid approach

The proposed method works well compared to existing methods. With further enhancements and structural adjustments, summaries generated with the proposed system tested with attorneys who have not yet been for real-time case. Since it is an unsupervised method that involves grouping, k-means and extracting the best-ranked proposition from each group and has computational advantages, it provides a promising start towards developing a fully functional Legal Case Summarizer [19].

2.7.5 An Approach to Automatic Text Summarization Using Simplified Lesk Algorithm and Wordnet

The proposed method uses unsupervised learning to summarize an input text according to a given percentage. First, apply the simplified Lesk method to each sentence to determine the weight of each sentence. Then the resulting sets of weights are arranged in descending order of their weights. The proposed method is based on semantic information extracted from the text. Therefore, various parameters such as formatting and the position of different places in the text are not taken into algorithm [20].

The proposed technical reporting method was found to work well because this text contains a smaller number of named entities in a sentence increases the number of important words in a sentence. As the number of significant words in a sentence increases, there will be more bright

images that will need to be cut out along with the text. As a result, the group weight is estimated more efficiently.

2.7.6 Learning Sentence Embeddings for Coherence Modelling and Beyond

This research is not completely related to the summarization of texts. However, when we learn abstract method, the sentences embedding is one of the most important areas. In this work, they demonstrated that a new method for capturing sentences learned through self-supervision can be applied effectively to text coherence problems [21].

To improve coherence modeling techniques, they introduced a self-supervision method called PPD to study sentence integration, which is based on the relationship between the meaning of a sentence and its position in the text. They used a recurring neural network to implement a new sentence insertion technique that is trained to assign a sentence to a discrete distribution that indicates.

2.7.7 Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

The author explained a focused neural network with encoder/decoder to interpret paragraphs of abstract text and shows that they achieve the best performance on two different corpora. It is also recommended to use some new models to solve critical summarization problems, and these problems are incorrectly modeled by the infrastructure, such as modeling key-words, capturing the hierarchy of sentence-to-word structure, and emitting words that are rare or unseen at training time [22].

In the context of the sequence-to-sequence model, we throw the mapping of the word input sequence in the source document into the target word sequence, which is called a summary. The author mentioned that a deep learning model (called sequence-to-sequence model) that maps one input sequence to another output sequence has successfully solved many problems, and recommend for machine translation.

2.7.8 An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation

Since transformers promise more differentiated modeling than repetitive units, the author explored these architectures for abstract summary tasks. Author also explained with the transformers without input sentence by trimming. Since the ground truth used in this field is used in training, and the generated words, regardless of whether they are correctly predicted or not, the author

proposes an improved structure to ERNIE-GEN for multi-flow training from sequence to sequence. It has a well-designed multi-flow attention architecture based on Transformer. This work is accompanied by the following estimates [24].

- **Infilling generation -** Instead of using the last ground truth word or last generated word in inference, but use the inserted artificial [ATTN] symbol and its location to gather insights into historical context.
- **Noise-Aware generation -** It is an effective method to corrupt the input sequence by randomly replacing words with arbitrary words in the vocabulary. It can alert the model of learning errors so that the model can recognize errors and ignore them during inference.

2.7.9 Abstractive Text Summarization Using Transformers

This article is an in-depth elucidation of the Transformer model from the well-known paper "Attention is all you need" by Google Research. In here author explained Google's Transformer model and implementation. Also, in this study they have explained the different between transformers and the sequence-to-sequence methods and both drawbacks [25].

The Transformer follows this general architecture and has a stacked self-attention layer and a point-to-point connection layer. These layers are fully connected to the encoder and decoder, and are shown in the left and right half of the architecture, respectively. The Transformer uses a self-attention mechanism in which all words in the input sequence are used to calculate attention weights at the same time, which helps parallelization. In addition, since the operation of each station in the transformer belongs to the same sequence of words, the complexity will not exceed. Therefore, the transformer proved to be an effective model and at the same time effective in terms of calculations. The author also explained the challenges faced by the RNN-based sequence about the transduction tasks and how the transformer model solves these challenges.

2.8 Chapter summary

The section studied several subjects appropriate in text summarization to classify numerous important results. The work inaugurates with text summarization methods. Two foremost methods for text summarization and compared those.

CHAPTER 3
PROBLEM ANALYSIS & METHODOLOGY

This chapter examines the problem in detailed and the procedure to accepted out the study on Sinhala text summarization on gazettes. It includes the prototype features, the way of abstractive and extractive methods to get the final outcome and limitation as well.

3.1 Sri Lanka Gazette Structure

In order to the summarization, first we have to identified the basic structure of the Government Gazette. The Sri Lanka Gazette (ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය) is published in Sinhalese, Tamil, and English which are three official languages of Sri Lanka. It publishes a few types for announced by various government departments [5]. Basically, it contains with six parts and a few numbers of sub sectors [6].

- **Part 1**
    - Government Notifications
    - Price Control Orders, Central Bank Notices
    - Posts (Vacant, Examinations, Results of Examinations)
    - Notices calling for Tenders
- **Part 2**
    - Laws, Supreme Court Notices
- **Part 3**
    - Land Settlement Notices
    - Final Orders
- **Part 4**
    - Proclamations and Appointments by the Governors
    - Provincial Councils Notifications
- **Part 5**
    - Books printed and registered under the printers and publisher's ordinance
- **Part 6**
    - List of Jurors

1. Gazette ID (Mandatory)
2. Gazette Published Date (Mandatory)
3. Part Number (Mandatory)
4. Section Details (Mandatory)
5. Type of Notice (Mandatory)
6. Sub-type of Notice
7. Act(s)
8. Title (Mandatory)
9. Notice (Mandatory)
10. Issued Date (Mandatory)
11. Issued By (Mandatory)

The details for the below format are similar for all the notices and few of them can be change due different type for notice formats. As an example: Application form type for vacancy notices, For the proposed project these won't be consider as valid documents and will be consider them on future enhancements.

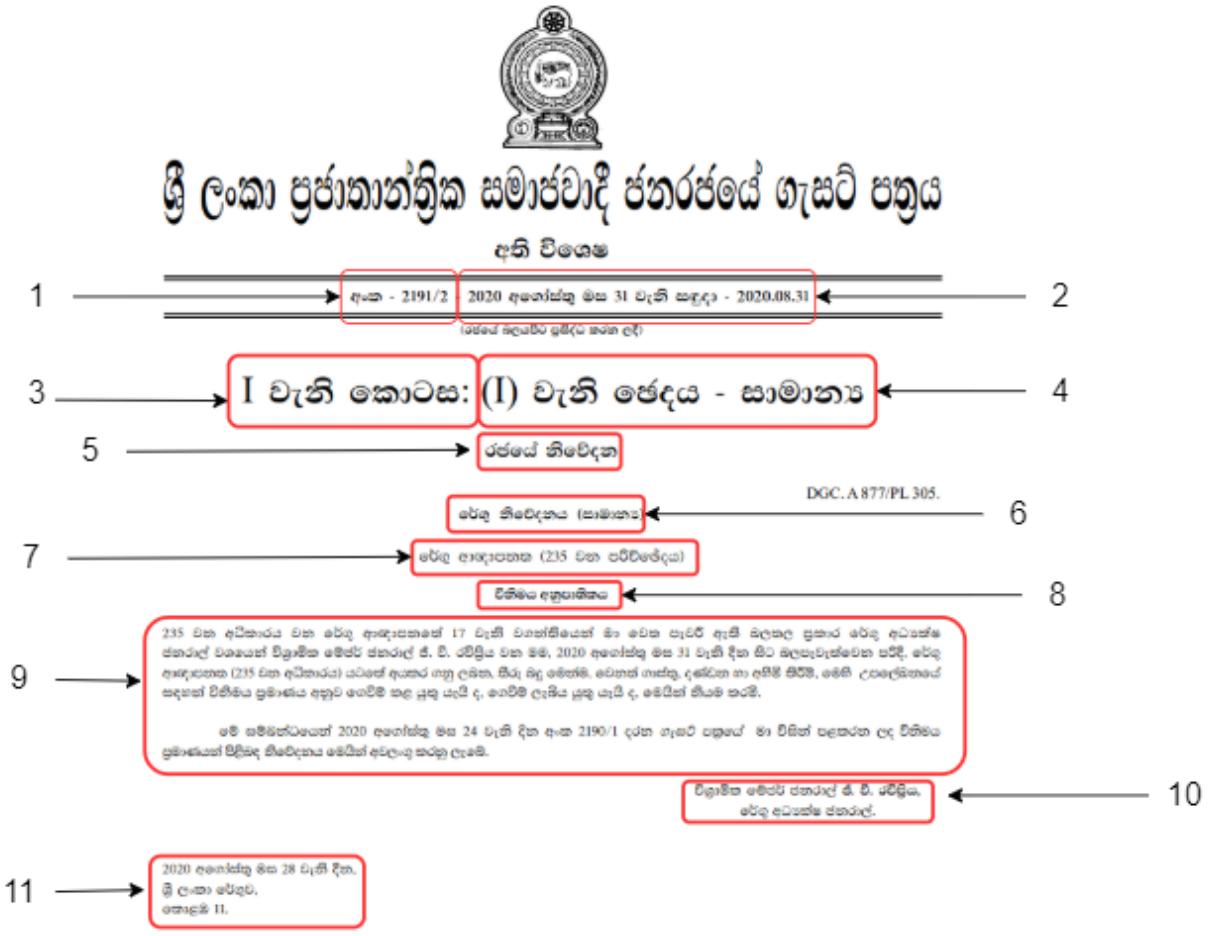


Figure 1 : Sample Gazette Notice

3.2 Features of Prototype

- **Identify correct type of Gazettes which have a long context.**

  As mentioned in the "1.7 Project Scope" section, the proposed project will only generate generic gazettes that satisfy a basic structure. If the uploaded PDF has the required structure, then this module will check the "notice" has a reasonable number of lines (not sentences) for summarization. If the notice does not have more than three lines the system will reject only the extraction summarization. Also, it will tokenize all the sentences and verify whether the document has a large context which need to be summarized.

- **Identify the important sentences in the source text and summarized the content.**

  1. **Scoring the sentence:** Select correct keywords, evaluate and select the highest scored sentences or words for the correct section.
  2. **Abstractive Summarization:** System will validate the template with the filled details.
  3. **Extraction Module:** From this validation, it will evaluate the sentences by removing any repeat words, adjectives, and stop words. Also, will validate the root from stemming words.
  4. **Identify Syntax Module:** Identify the sentences by using rules and check and verify alternatives and select the best syntax format for each sentence.
  5. **Compression and Concur Module:** Evaluate the concur words and combination sentences were properly implemented.

- **Organized the details and output the final outcome.**

  The final outcome of the summarization will be shown in template-based approach and here is a sample UI temple. "Summery of the notice" will contain the extractive text summery of the gazette notice. All the Acts can be found in the "Related Acts" section

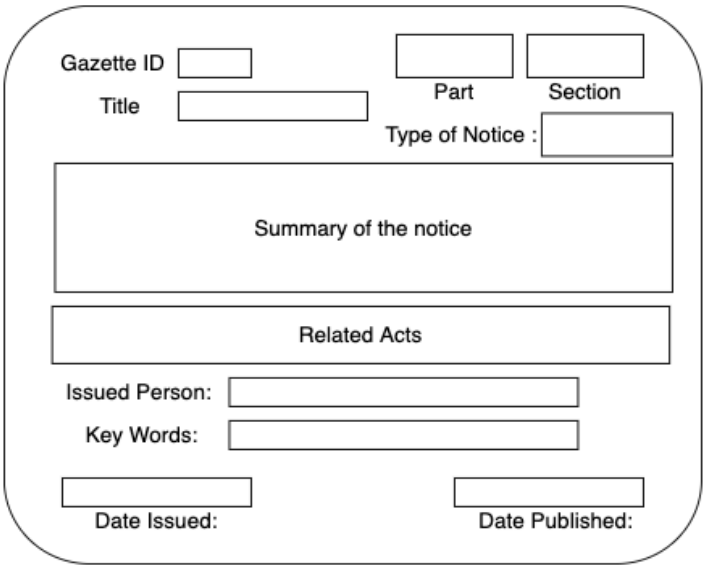


Figure 2 : UI Wireframe

3.3 Limitations

Since inadequate resources and studies supported for Sinhala language, this Sinhala summarization become complex and a real challenge to fulfil the scope. Since there are no define solid infrastructure and WordNet for Sinhala language this might be the biggest limitation in the research. Hence, we are following few methods defined for rich languages such as English, Chinese and Hindi text summarization for build a valid platform for this research.

Consequently, the scope of the project does not include the final result in the form of Sinhala written grammar (ex: මම → මී, අපි → මු, නුඹ → හි). Although, as mentioned in the above we are not focusing on all type of gazettes since some of them are having complex type of notices (such as sample application for vacancies). As a beginning and early phase in this research, it will be converging on single document and common type of gazettes which will fulfil the requirement.

The proposed solution will take the Gazette input as PDF version. The PDF file to convert to text document is another huge challenge and final output may limit to that conversion. Finally current solution will exclude exceptional gazettes such as two columns' notices, multiple issues, multiple notices and any complex gazettes.

3.4 Methodology

In this context, it will explain the methodology that will carry out the research on Sinhala text summarization. This research mainly focusses on abstractive method summarization (templated based approach) and it also interacts with the extractive approach for summarize the notice description.

3.4.1 Tokenized the sentences

Tokenization is the process by which big quantity of text is divided into smaller parts called tokens. Natural language processing is used for construction applications such as Text organization, intellectual chatbot, sentimental analysis, language translation, etc. It becomes vital to understand the pattern in the text to achieve the above-stated purpose. These tokens are very useful for discovery such patterns as well as is measured as a base step for stemming and lemmatization.

Table 1 : Example of tokenized sentences

ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය
අති විශේෂ
අංක - 219/2 - 2020 අගෝස්තු මස 31 වැනි සඳුදා - 2020.08.31
(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)
I වැනි කොටස: (I) වැනි ඡේදය - සාමාන්‍ය
රජයේ නිවේදන
……..
විශ්‍රාමික මේජර් ජනරාල් ජී. වී. රවිප්‍රිය,
රේගු අධ්‍යක්ෂ ජනරාල්.
2020 අගෝස්තු මස 28 වැනි දින,
ශ්‍රී ලංකා රේගුව,
කොළඹ 11.

3.4.1.1 Remove Punctual/ Special characters

This is similar to word tokenization and identify and remove unwanted characters from the sentence(s).

Table 2 : Tokenized sentences after format

ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය
අති විශේෂ
අංක
219/2
2020 අගෝස්තු මස 31 වැනි සඳුදා
2020.08.31
රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී
I වැනි කොටස
I වැනි ඡේදය
සාමාන්‍ය
රජයේ නිවේදන
……..

3.4.1.2 Identifying the Keywords

Article keywords are primarily determined based on how often the term is used. If a term appears more frequently in a document and in most corpus documents, then the term is considered less important. When calculating the frequency of terms in each document of the corpus, the functional words defined in Corpus-based Sinhala Lexicon [26] are omitted. They identified 440 words, including the word Sinhala (නිත්‍ය පද), among conjunctions, modifiers, interjections, particles, and postpositions.

3.4.2 Scoring the sentence

Assign weights on sentences in the given document. For that we have to classified the selected words.

- **Word Type:** Identify the words according to the nouns, pronouns, verbs, adjectives, adverbs, conjunctions, prepositions, and interjections. Store them in separately.
- **Cue words:** Cue words are connective expressions that link spans of discourse and signals semantic relations in a text. [16]
- **Title Words:** The words appear in the title, subtitles and headings are considered as title words. [16]
- **Key Words:** The list of keywords which is retrieved by 3.4.1.2 can be identified by sorting the list of words in the document by their frequency to score the sentences.
- **Word Location:** Word location is the feature to assign a weight for a sentence based on its position in the document. [16]

To weight the sentences, we have to follow an approach and at the moment we have proposed the following scoring method. If the particular keyword is available in sentence, then, that sentence will be categorized into the above sections.

Table 3 : Sample keyword list according to the section

Keywords	Section
මස, ජනවාරි, පෙබරවාරි, …, දෙසැම්බර්	Date issued and Date published
වැනි, කොටස,	Part Number
නිවේදන, දැන්වීම්, නිවේදනය	Title and Sub title
වගන්තිය, ආඥාපනත, නීතිය	Related Acts
නීතිඥ, ජනරාල්, කොමසාරිස්, නිලදාරී	Issued by

Here we are using the word location or the position identified the correct details for above sections.

Table 4 : Word location with keywords

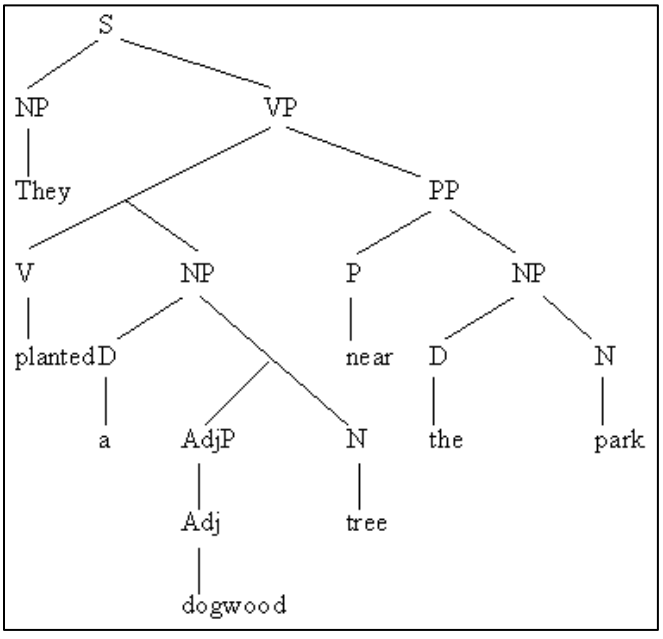
Keywords	Prefix of suffix word	Section
ජනවාරි,පෙබරවාරි, …	මස	Dates
වැනි	කොටස	Part Number
වැනි	ඡේදය	Sub-part
වගන්තිය, ආඥාපනත, නීතිය	<Number, Letter>	Related Acts
නීතිඥ, ජනරාල්, කොමසාරිස්, නිලදාරී	අධ්‍යක්ෂ, තේරීම්භාර	Issued by

3.4.3 Stemming words using Data

Stemming is an important process in the field of NLP. This is the process of reducing the variable form of a word to its root. Many NLP applications that use words as components use roots to extract the roots of words. This is a very efficient and lightweight approach compared to morphological analysis. Languages like English have a high degree of lineage, but the algorithms they use don't work well in prone languages like Sinhalese. Sinhalese is a very refracted language, so there are many forms of the same concept. This situation is strongly influenced by the frequency of the term, so you must stop the word before defining it. These are not previous attempts in the literature to determine the correct highlighting algorithm for Sinhala. Therefore, to determine the root of each word, we define two stem extraction algorithms as described below and choose an approach.

3.4.4 Identify Syntax Module

One of the most important steps in the above four steps are to define the grammar and identify the correct syntax and semantic rules for a sentence. We have to define rules to identified the proper structure for the words since the message is in grammar format. Below image is a sample structure for English language.



- **S → NP, VP**
- **NP → N**
- **NP → Det, Adj, N**
- **NP → Det, N**
- **VP → V, NP, PP**
- **PP → P, NP**

Figure 3 : Simple Syntax Module for English

3.4.5 Finalized the extractive summary

For this, we need to extract the keywords and the sentences which has low scoring rate compare to the others.

3.4.5.1 Removal Module

We have to identified the sentences which are no need for the core detail of the context and remove them. For others words we have to design a summarizing set of syntactic rules including the removal of:

- Appositive noun phrases
- Relative clauses and adverbial clauses
- Adjective phrases and adverbial phrases
- Prepositional phrases
- Content within parentheses and other parentheticals.

3.4.5.2 Compression and Concur Module

After selecting the sentences, the text compression module evaluates decides whether to remove certain phrases or words in the selected sentences according to the weight score. Sample output for above example will be shown as below.

Table 5 : Comparison of Notice and final outcome

Complete Notice	Extractive Summary
235 වන අධිකාරය වන රේගු ආඥාපනතේ 17 වැනි වගන්තියෙන් මා වෙත පැවරී ඇති බලතල ප්‍රකාර රේගු අධ්‍යක්ෂ ජෙනරාල් වශයෙන් විශ්‍රාමික මේජර් ජනරාල් ජී. වී. රවිප්‍රිය වන මම, 2020 අගෝස්තු මස 31වැනි දින සිට බලපැවැත්වෙන පරිදි, රේගු ආඥාපනත (235 වන අධිකාරය) යටතේ අයකර ගනු ලබන, තීරු බදු මෙන්ම, වෙනත් ගාස්තු, දණ්ඩන හා අහිමි කිරීම්, මෙහි උපලේඛනයේ සඳහන් විනිමය ප්‍රමාණය අනුව ගෙවීම් කල යුතු යැයි, ගෙවීම් ලැබිය යුතු යැයි ද, මෙයින් නියම කරමි.	2020 අගෝස්තු මස 31 වැනි දින සිට අයකර ගනු ලබන තීරු බදු, වෙනත් ගාස්තු, දණ්ඩන හා අහිමි කිරීම්, මෙහි සඳහන් විනිමය ප්‍රමාණය අනුව ගෙවීම් කල යුතු යැයි, ගෙවීම් ලැබිය යුතු යැයි, නියම කරමි.

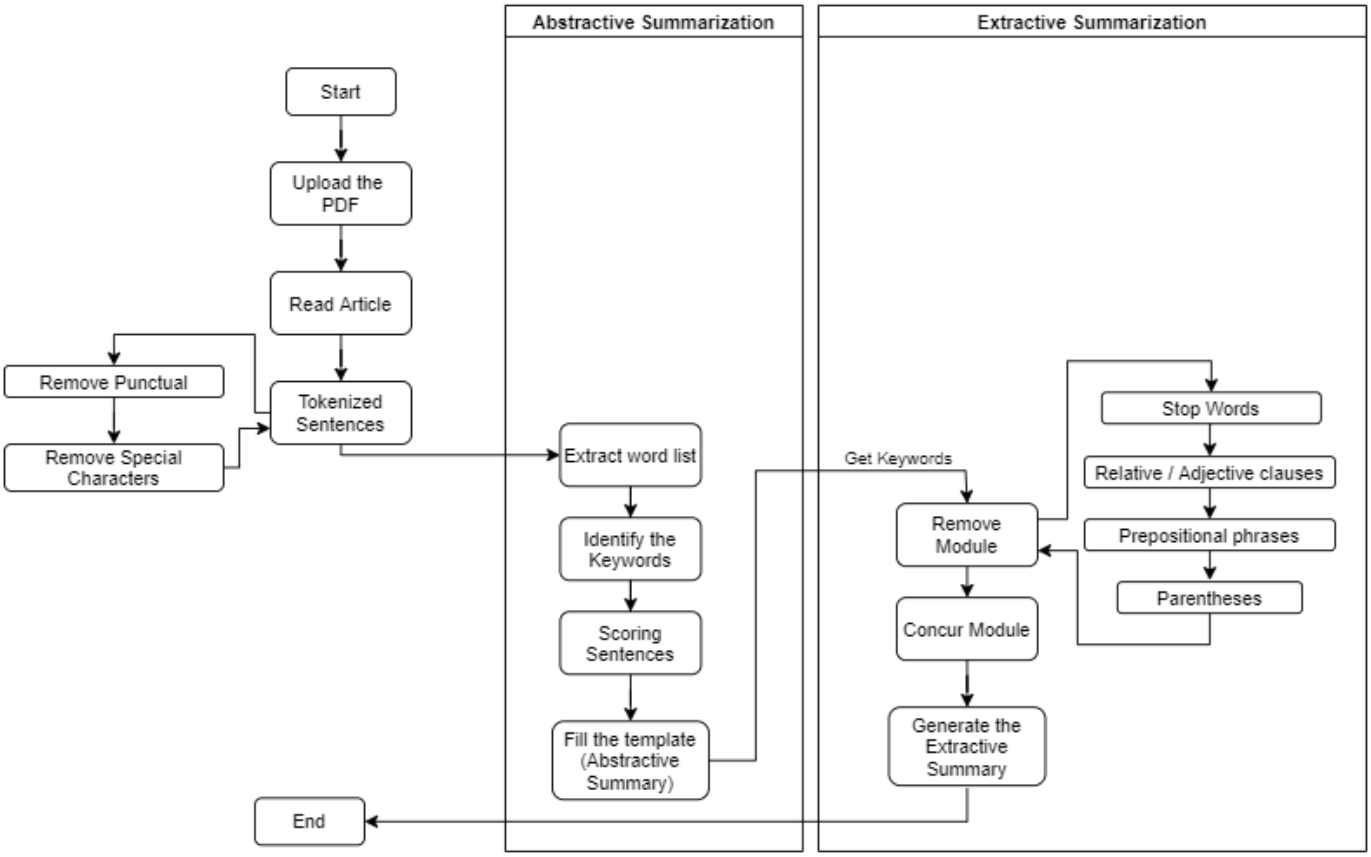


Figure 4 : Proposed Solution (High-level algorithm)

According to above figure 4 system will take the input keywords for extractive summary process, from the output of abstractive summary details. The reason for that is it will help to identify any important keywords or repetitive keywords in the notice body.

3.5 Chapter summary

The chapter examined the Sri Lanka gazette structure and how the solution will cater the final outcome according to that. The study commences the process of the system works according to the techniques defined in this chapter. Also, explained briefly about limitation which will needed for next chapter.

CHAPTER 4
EVALUATION, EXPERIMENT AND TEST RESULTS

In this chapter, the implemented system is evaluated in terms of technology, ease of use, and concept. Furthermore, the experiment was carried out using the method developed for the selected dataset. The results of these experiments and their evaluations are also described with corresponding assumptions.

4.1 Validating the PDF or the Gazette

User can upload the document to the system through the URL. The proposed system is evaluating this document as the first step. Here are the criteria that the system is validating.

- URL is Accessible
- URL contain a readable PDF document
- PDF document is a Sinhala Gazette
- PDF has only one page (refer to the below assumption)
- Content should be in one text column
    - Due to the limitation of PDF readers, multi text column notices are hard to identify exact lines with the content. This can be done by complex algorithm but since this is not related to this project scope, multi text column notices will be considered in future improvements.

Assumption:

If the gazette has more one-page, extra page might be including tables or forms which is complex to summarized at this level. Hence if the PDF has two or more page, we are not using those as the dataset.

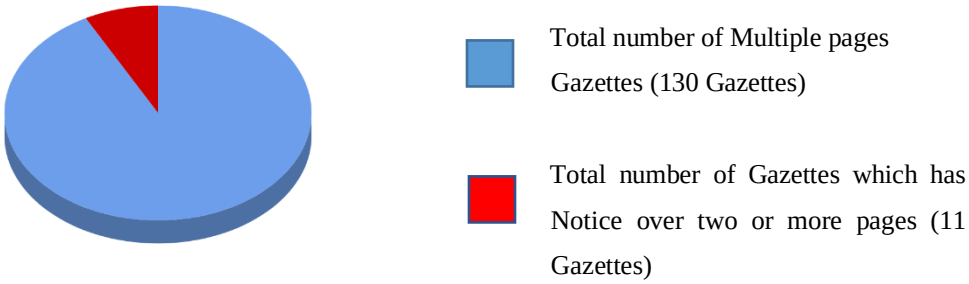


Figure 5 - Comparison of gazettes which has notice over two pages with gazettes which has more than one pages

4.2 Data Set

To study the automatic summarization of Sri Lankan Sinhala text, we decided to research well-structured Sri Lankan articles in a specific field which is the Government Gazettes. After reviewing many regular newsletters, we have determined the following.

- Department wise and issuer wise, basic template of the gazette is differing
- Notice barley have multiple paragraphs. We have ignored number of paragraphs from the evaluation
- The average number of sentences per notice is two
- The average number of words per sentence is 24 which is quite long compare to other articles in Sinhala
- The average number of words per gazettes' notice is 57 words which makes stemming is quite complex process to implement
- Exceptional cases such as tables, multiple issuers and different structure is making difficult to create a standard way of identifying the context

We have collected over 550 gazettes from the Department of Government Printing website. Out of those gazettes 450 gazettes were elected as the correct data set for this project. It was assumed that over 17000 words is to be sufficient to represent the language for research purposes, such as automatically text summarization in this domain. The collected gazettes and data are stored in computer as PDF and TXT file formats and table 4.1 shows the basic statistics of the defined dataset.

Table 6 : Basic statistics of Data set

Number of Gazettes	450
Total number of sentences	856
Total Words	17656
Average number of sentences per gazette	2
Average number of words per Gazette	57
Average number of words per Sentence	24
Total number of distinct words	9384
Average distinct words per Gazette (without Common words)	21

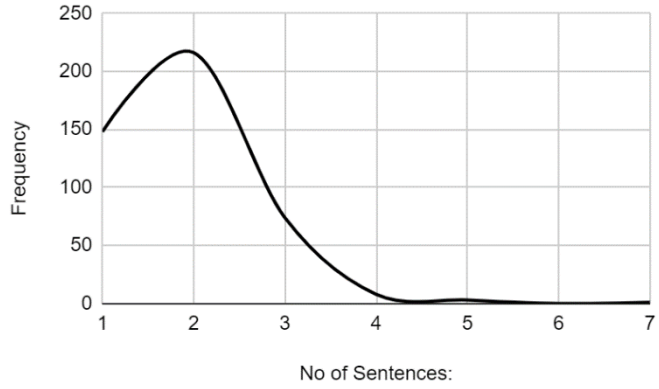


Figure 6 - Number of sentences per Gazettes vs Frequency

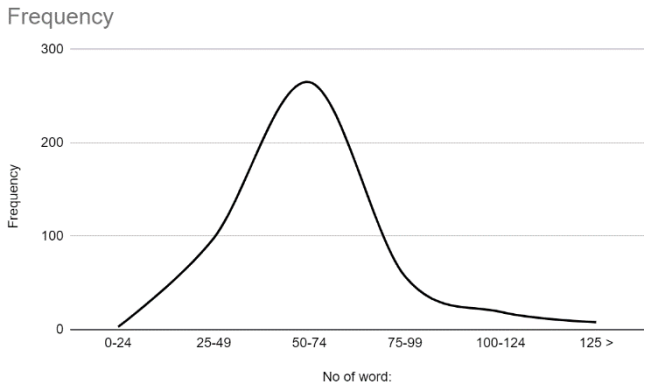


Figure 7- Number of words in Gazette vs Frequency

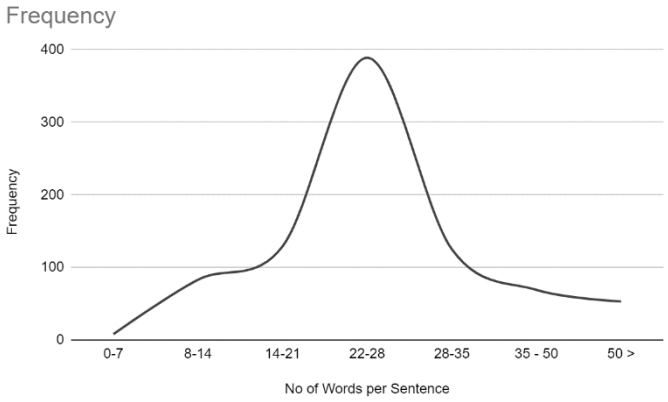


Figure 8 - Number of words in sentence vs Frequency

4.3 Evaluating Summaries by Manual

The suggested automatic text summarization Sinhala text annotations will be evaluated on 50 manually selected gazettes from 450 dataset mentioned above. However, due to lack of human resources, the number of manually summarized documents were limited to this number. It was found that almost 10% of the manual tests are sufficient to assess the effectiveness of the summering process. Following evaluation criteria based on this 50 gazettes and final results attached in the Appendix B section.

4.4 Defining the Evaluation Criteria

Abstractive Summarization:

In order to assess the quality of the computer-extracted abstracts compared to the manual summaries, the accuracy and responsiveness of the computer-generated abstracts were calculated. Calculating accuracy and retrieval rate to measure the correlation between machine-generated datasets and manual datasets is a mature technology, specifically in the field of outline Precision and Recall. Precision is well-defined as the proportion of related examples retrieved and Recall is defined as the proportion of related instances retrieved. [23].

Precision: The number of precise sections divided by the amount of outcome forecast. Here we have defined the template parts as sections.

$$\frac{Number\ of\ Correct\ Sections}{Predected\ Correct\ Sections}$$

Recall: It is the number of precise positive outcomes alienated by the amount of all precise trials. Here we have defined the template parts as sections.

$$\frac{Number\ of\ Correct\ Sections}{Correct\ Sections + Unpredected\ correct\ Sections}$$

Giving to the above calculations, if it attempts to increase the recall rate by retrieving more instances, it will cause to decrease the Precision rate. To get the maximum values for both

Precision and recall, F-Score is calculated. F-Score reaches its best value at 1 and worst score at 0. [23]

$$F\ Score = 2\ \text{x}\ \frac{\text{Precision}\ \text{x}\ \text{Recall}}{\text{Precision}\ +\ \text{Recall}}$$

This F-Score measure was calculated for each computer generated and manually extracted summaries to evaluate the performance of the proposed methodologies. [23]

Extractive Summarization:

Evaluate Words:

In view of the scoring criteria of the extractive summarization, we used methods to score the keywords and final results. The built-in vector word representation is a representation of a set of word vectors generated by an embedded method (such as Word2Vec or GloVe) for some intermediate subtasks. These subproblems are generally quick and easy to calculate, thus allowing you to understand the system used to generate word vectors. The intrinsic evaluation should usually return a number to indicate the performance of these word vectors on the evaluation subtask.

- Takes every word in the notice as inputs
- Clustering common words into different groups. Examples:
    - අනිවානර්යෙන්, වහාම, අණ... → Critical words
    - කාල, වකවානුව, දින… → Time related words
    - දක්වා, තුරා, සිට → Duration related words

After the computerization grouping, evaluation criteria which are defined above can be used for this clustering and evaluated the results. F-Score value will be considered from these two outputs.

$$Precision\ =\ \frac{Correct\ words\ in\ each\ cluster}{Expected\ words\ in\ particular\ cluster}$$

$$Recall\ =\ \frac{Correct\ words\ in\ each\ cluster}{All\ the\ words\ selected\ by\ machine\ for\ particular\ cluster}$$

We also use ROUGE, which stands Recall-Oriented Understudy for Gusting Evaluation. Basically, it is a set of indicators used to assess automatic text summarization and automatic translation. It works by comparing an automatically produced summary or translation against a set of manual summaries. ROUGE results will evaluate extractive summary algorithm accuracy. [1]

ROUGE-N and ROUGE-S can be considered as the text granularity of the comparison between the system summary and the manual summary. ROUGE-N overlap of N-grams between the system and reference summaries, ROUGE-S Skip-bigram is any pair of words and based co-occurrence statistics. Furthermore ROUGE-N can be divided into two sub categories which are, ROUGE-1 refers to overlap of unigrams between the system summary and reference summary and ROUGE-2 refers to the overlap of bigrams between the system and reference summaries. (O - Original Document, HS - Human Summary, MS - Machine Summary)

$$Words_{(O-HS)} = \frac{\text{Number of words in the Manunal Summary}}{\text{Number of words in Gazette Notice}}$$

$$Words_{(O-MS)} = \frac{\text{Number of words in the Machine Summary}}{\text{Number of words in Gazette Notice}}$$

$$ROUGE = \frac{\text{Number of words in the Machine Summary}}{\text{Number of words in the Manunal Summary}}$$

Weighting for Keywords

In the manual approach 50 summaries were created only using the keyword feature based on most weighted words of each gazette. The original words of the gazette title were retained to maintain the flow of information of the extracted summary. Then the Precision, Recall and then the F-Score for each 50 articles were calculated with respect to their corresponding manually extracted summaries.

$$Precision = \frac{Correct\ Keywords\ Selected\ by\ Machine}{Expected\ Keywors}$$

$$Recall = \frac{Correct\ Keywords\ Selected\ by\ Machine}{All\ Keywords\ Selected\ by\ Machine}$$

Table 4.2 shows a sample of different F-Score values calculated based on selected 10 gazettes. Approach used to identify each of the above features were evaluated based on the F-Score values generated by comparing machine extracted summaries against the human extracted summaries.

Table 7 : Keyword Weighting Results for 10 gazettes

Sample File	Precision	Recall	F-Score
2189-48_S.pdf	0.48	0.84	0.61
2196-22_S.pdf	0.85	0.17	0.28
2181-24_S.pdf	0.56	0.25	0.34
2228-41_S.pdf	0.69	0.27	0.39
2193-31_S.pdf	0.72	0.44	0.55
2196-06_S.pdf	0.78	0.74	0.76
2196-43_S.pdf	0.36	0.53	0.43
2217-39_S.pdf	0.19	0.17	0.18
2188-29_S.pdf	0.92	0.71	0.80
2182-32_S.pdf	0.50	0.72	0.59

Table 4.3 contain full results of above sample 10 data sets.

Table 8 : Final F-Score results for 10 Gazettes

No	Sample File	No of Words	Abstractive Summarization	Extractive Summarization		Average F-Score
				Words	Keywords	
1	2189-48_S.pdf	73	0.91	0.39	0.61	0.64
2	2196-22_S.pdf	108	0.88	0.42	0.28	0.53
3	2181-24_S.pdf	91	0.82	0.48	0.34	0.55
4	2228-41_S.pdf	76	0.82	0.55	0.39	0.59
5	2193-31_S.pdf	58	0.85	0.41	0.55	0.60
6	2196-06_S.pdf	65	0.83	0.53	0.76	0.71
7	2196-43_S.pdf	98	0.81	0.18	0.43	0.47

8	2217-39_S.pdf	165	0.69	0.24	0.18	0.37
9	2188-29_S.pdf	52	0.88	0.57	0.8	0.75
10	2182-32_S.pdf	78	0.84	0.48	0.59	0.64

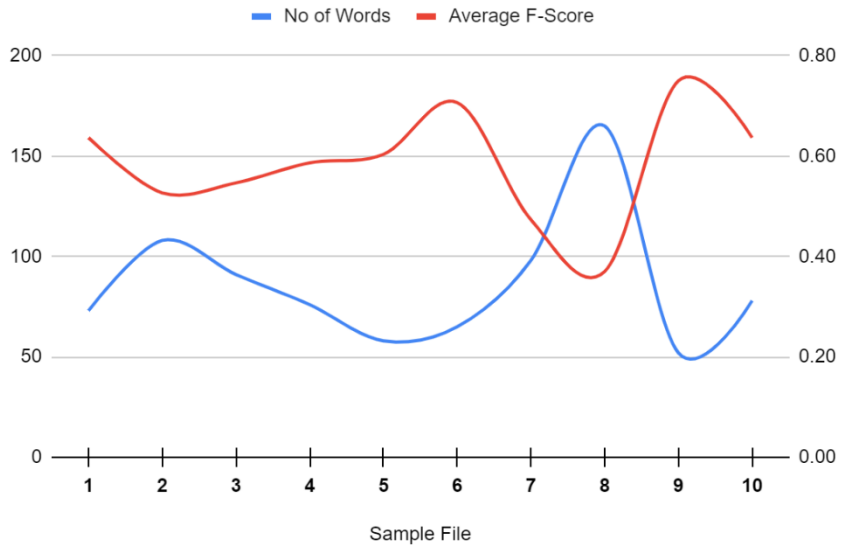


Figure 9 - Average F-Score vs No of Words in the Gazette

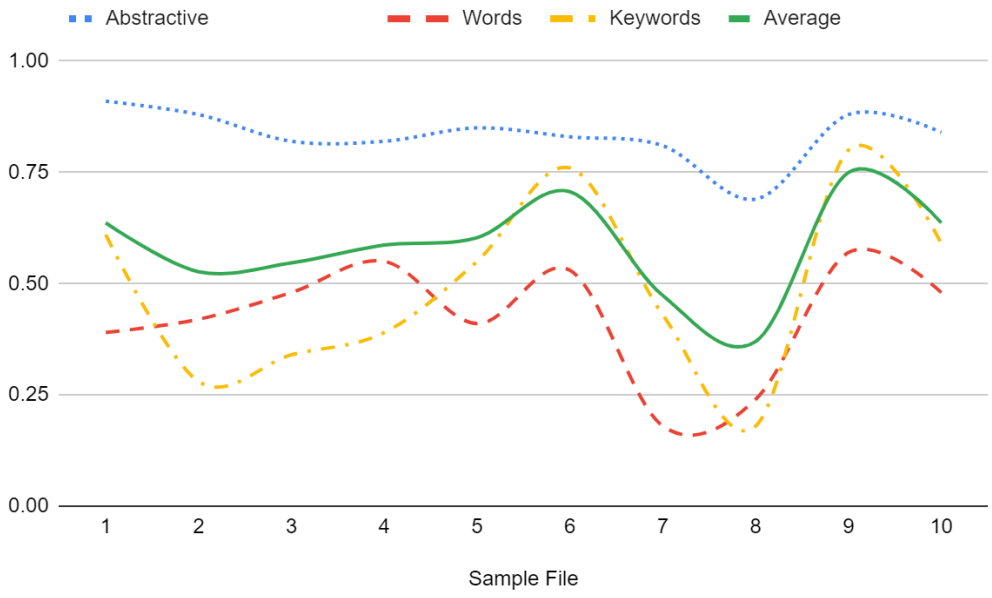


Figure 10 - Behavior of Abstractive, Extractive (words, keywords) F-Score

4.5 Experiments with the Test Data

The full results of the data set mentioned in table 4.1 are shown in the table 4.4. The 450 results values were evaluated by manually by not going through the original gazettes.

Table 9 : Abstract Summarization results for 450 gazettes

Section	Pass	Failed	Percentage
Gazette No	445	5	99%
Title	432	18	96%
Part and Section	440	10	98%
Issued Date	448	2	99%
Acts	419	31	93%
Issuer	405	45	90%
Issued Department	371	79	82%

In the below table 4.5 we have check the following formular and categorized the extractive summarization for gazette notice.

$$Words_{(O-MS)} = \frac{\text{Number of words in the Machine Summary}}{\text{Number of words in Gazette Notice}}$$

$Words_{(O-MS)}$ cannot be less than or equal zero and cannot be larger than one. If the value is exactly one, then we can assume that original notice was selected as the summarized output. Assumption is summarized output should be 1/3 no of words compare to the original notice.

Table 10 : Extractive Summarization results for 450 gazettes

$Words_{(O-MS)}$ Range	No of Gazettes	No of Gazettes
0 - 0.33 (Good)	97	22%
0.34 - 0.66 (Medium)	238	53%
0.67 - 1 (Worst)	115	26%

Figure 7 shows the implemented solution performance against the number of documents at a single process. This process executed couple of times and average value taken as the final results for the below chart. Tested machine had 16 GB RAM and 2.2 GHz Quad-Core - i7 CPU.

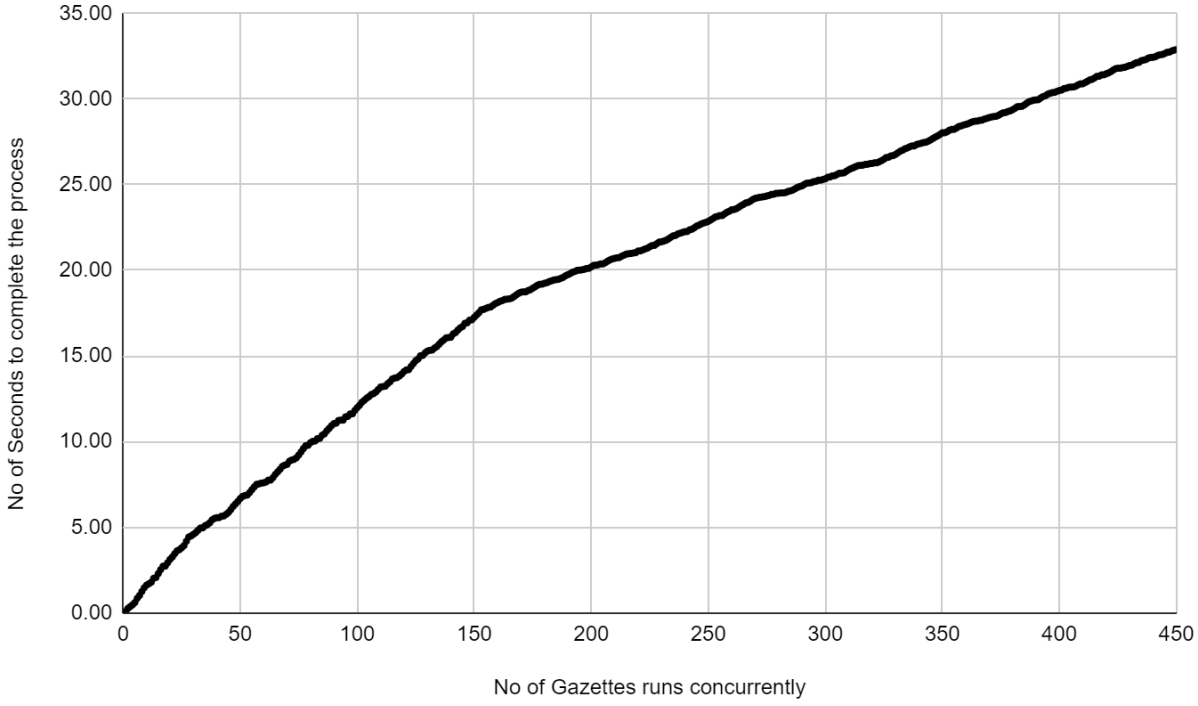


Figure 11- Summarization Performance

4.6 Conclusion

Comparing both abstractive and extractive summarization results, the algorithm for extractive summary needs to be improve to overcome high accuracy output. Somehow the logics and the methodology for this notice summarization needs be re-structure to fulfil the current limitations.

CHAPTER 5
CONCLUTION AND FUTURE ENHANCEMENTS

This chapter summarizes the proposed research results, the results obtained, and future work that can be done to improve the quality of summarization designed for Sri Lankan gazettes. It will deliberate the challenging and encounters met throughout the project progress phase.

5.1 Challenges and Learning Outcomes

Most critical challenge that faced during this project is to implement a generalized solution. Due to a lot of gazettes formats and context issue currents system has many limitations in this phase. Not like other Sinhala document domain, gazettes do not have many sentences but have large no of words for each sentence. Because of that we could not able to apply all most all the approaches which are already researched and developed for Sinhala language.

Apart from that, time management, lack of human and research resources are a few of other changelings that faced during this project. Even though there a few of researches available in the Sinhala language, those were help a vast of amount for this project.

As learning outcome, this project help to improve the knowledge of how the automatic text summarization is working, what are the approaches, pros and cons, what are difficulties that we can find in Sinhala language and Sinhala summarization. The experience gathered from this project is a huge turning point for me when comes to the NLP and text summarizations which are interesting research domains in modern days.

5.2 Future Enhancements

This research was carried out based approaches stated in chapter 2 for automatic text summarization. The foremost objective on choosing methods is to transmit the study with lowest accessible language possessions for Sinhala language. Future extensions of this research can be carried out in many directions and this section is intended to describe some of these in details.

5.2.1 Reduce application limitations

As explained in the chapter 4, the project is having few limitations due Java PDF reader and due to some exceptional gazettes. These limitations must be fulfilling in future to overcome and solve the domain issues which explained in the chapter 1. Mostly, we have to improve the code base and research on more exceptional gazette and needed built and generic solution.

5.2.2 Improve keyword areas and test on more gazettes

Since the abstraction summarization is highly based on keyword knowledge, we have to increase the keywords area. That will increase the efficiency of the final outcome. Also, by testing the solution on more gazettes by increasing the dataset, we can easily increase the knowledge area as well.

5.2.3 Support Multi Pages gazettes

The biggest limitation on this solution is, not handling the multiple pages. Out of the large number of gazettes there are significant number of gazettes which has multiple pages. This will be kind of deeper research area in this gazette summarization domain. We have to identify context, how the context is behaving, identify forms, tables, multiple notices and a way to summarize those are quite challengeable and need deep analysis.

5.2.4 Optimize the logics and performances

Logic optimization is a needful for every solution. We have to monitor how the system is performance and the accuracy of the outcome when there are multi pages or large context.

5.2.5 Store keywords in to a database

In this phase we have integrate the solution to the database. Also, we generate the summary of the gazette by analyzing particular gazette at that moment. But if we can store the keywords

and store the outcome of one gazette and use that knowledge to analyzes the next gazette or similar gazette that will make a significant improvement on efficiency of final outcome.

5.2.6 Make the solution open source give code access to the public

We believe this solution will solves a major problem in Gazette domain. Also, project is based on Sinhala text summarization which is a rising research area, if someone interested in that domain, they can use our keywords classification and logics to make better solution or improve this solution in to a better approach.

5.3 Conclusion

As mentioned in Chapter 1, no attempt has been made to summarize the Sinhala gazettes. However, various types of research have been conducted on Sinhala and Sinhala summarizations. Many different methods have been tested against other languages (such as English) to determine the best way to automatically generalize the summaries for large contexts.

This study focuses primarily on finding the best way to automatically summarize Sri Lankan gazettes, which have the fewest language resources. Therefore, the experiment is based on the classic method of automatic text composition. Experimental results show that the characteristics of several subjects identified by researchers for languages such as English can be used in Sri Lanka for the same purpose. This project showed that, implementing such solution is possible.

References

[1] - https://www.newworldencyclopedia.org/entry/Abstract_(summary) [Accessed: 28-NOV-2020]

[2] - Tldx, K. and Lu, D.F. (2017), "Neural Extractive Text Summarization with Syntactic" https://www.aclweb.org/anthology/D19-1324.pdf pp. 54-58 [Accessed: 23-AUG-2020]

[3] - Arukgoda, J. et al. (2015), "A Word Sense Disambiguation Technique for Sinhala". Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014.

[4] - https://en.wikipedia.org/wiki/Government_gazette [Accessed: 28-AUG-2020]

[5] - https://en.wikipedia.org/wiki/The_Sri_Lanka_Gazette [Accessed: 28-AUG-2020]

[6] - http://www.documents.gov.lk/si/gazette.php [Accessed: 28-AUG-2020]

[7] - Allahyari, M., Trippe, E.D. and Gutierrez, J.B. (2013) Text Summarization Techniques: A Brief Survey pp 21-23

[8] - Moratanch, N. (2016). A Survey on Abstractive Text Summarization pp 40

[9] - Nallapati, R. et al. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. Available from http://arxiv.org/abs/1602.06023 [Accessed: 31-JAN-2021]

[10] - Hingu, D., Shah, D. and Udmale, S.S. (2017). Automatic Text Summarization of Wikipedia Articles pp 2–4

[11] - Nouf Altmami, et al. (2020), "Automatic summarization of scientific articles: A survey", https://www.sciencedirect.com/science/article/pii/S1319157820303554 [Accessed: 30-AUG-2020]

[12] - W V Welgama (2012) "Automatic Text Summarization for Sinhala" pp. 11-29

[13] - Jiacheng Xu, Greg Durrett, "Neural Extractive Text Summarization with Syntactic Compression", pp. 3292-3295, https://www.aclweb.org/anthology/D19-1324.pdf [Accessed: 30-AUG-2020]

[14] - Liu, J. (2017). A Multi-Level Encoder for Text Summarization pp 2-8

[15] - Mihalcea, R. and Tarau, P. (1998). TextRank: Bringing Order into Texts pp 85, 91-92

[16] - Welgama, V. et al. (2011). Towards a Sinhala Wordnet. Preceedings of Conference on Human Language Technology for Development, HHLTD 2011, (May), 39–43.

[17] - Wijesiri, I. et al. (2014). Building a WordNet for Sinhala. GWC 2014: Proceedings of the 7th Global Wordnet Conference, (June)

[18] - Aysa Asa1, and et al. (2017) "A Comprehensive Survey on Extractive Text Summarization", pp. 226-230, http://www.ajer.org/papers/v6(01)/ZH0601226239.pdf [Accessed: 30-AUG-2020]

[19] - Varun Pandya (2019), "Automatic Text Summarization of Legal Cases: A Hybrid Approach", pp. 38-41 https://www.academia.edu/42950597/AUTOMATIC_TEXT_SUMMARIZATION_OF_LEGAL_CASES_A_HYBRID_APPROACH [Accessed: 31-AUG-2020]

[20] - https://www.academia.edu/40428790/An_Approach_To_Automatic_Text_Summarization_Using_Simplified_Lesk_Algorithm_And_Wordnet [Accessed: 28-AUG-2020]

[21] - Alok Ranjan Pal (2013), "An Approach to Automatic Text Summarization Using Simplified Lesk Algorithm and Wordnet", pp 15-21

[22] - Bowen Zhou, Ramesh Nallapati et al. (2016) "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond" pp. 281-282

[23] - O.S Wimalasuriya (2019), "Automatic Text Summarization for Sinhala", pp. 6-21

[24] - Dongling Xiao et al. (2020), "ERNIE-GEN: An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation" pp. 2-5

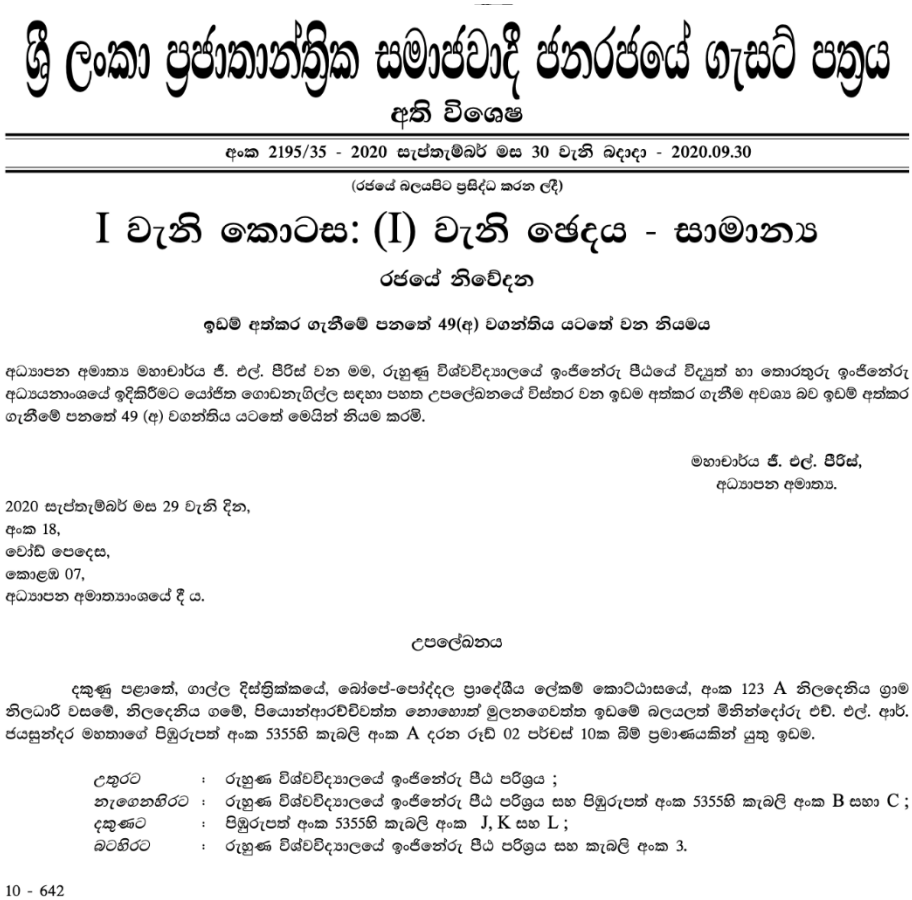
[25] - https://towardsdatascience.com/transformers-explained-65454c0f3fa7 [Accessed: 03-DEC-2020]

[26] - Ruvan Weerasinghe, Dulip Herath, Viraj Welgama (2009), "Corpus-based Sinhala Lexicon", pp 19-20

Appendix A

Samples of Source Article and Machine Extracted Summaries

Sample 1:



ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය

අති විශෙෂ

අංක 2195/35 - 2020 සැප්තැම්බර් මස 30 වැනි බදාදා - 2020.09.30

(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)

I වැනි කොටස: (I) වැනි ඡෙදය - සාමාන්‍ය

රජයේ නිවේදන

ඉඩම් අත්කර ගැනීමේ පනතේ 49(අ) වගන්තිය යටතේ වන නියමය

අධ්‍යාපන අමාත්‍ය මහාචාර්ය ජී. එල්. පීරිස් වන මම, රුහුණු විශ්වවිද්‍යාලයේ ඉංජිනේරු පීඨයේ විද්‍යුත් හා තොරතුරු ඉංජිනේරු අධ්‍යයනාංශයේ ඉදිකිරීමට යෝජිත ගොඩනැගිල්ල සඳහා පහත උපලේඛනයේ විස්තර වන ඉඩම අත්කර ගැනීම අවශ්‍ය බව ඉඩම් අත්කර ගැනීමේ පනතේ 49 (අ) වගන්තිය යටතේ මෙයින් නියම කරමි.

මහාචාර්ය ජී. එල්. පීරිස්,
අධ්‍යාපන අමාත්‍ය.

2020 සැප්තැම්බර් මස 29 වැනි දින,
අංක 18,
වෝඩ් පෙදෙස,
කොළඹ 07,
අධ්‍යාපන අමාත්‍යාංශයේ දී ය.

උපලේඛනය

දකුණු පළාතේ, ගාල්ල දිස්ත්‍රික්කයේ, බෝපේ-පෝද්දල ප්‍රාදේශීය ලේකම් කොට්ඨාසයේ, අංක 123 A නිලදෙනිය ග්‍රාම නිලධාරී වසමේ, නිලදෙනිය ගමේ, පියොන්ආරච්චිවත්ත නොහොත් මුලනගෙවත්ත ඉඩමේ බලයලත් මිනින්දෝරු එච්. එල්. ආර්. ජයසුන්දර මහතාගේ පිඹුරුපත් අංක 5355හි කැබලි අංක A දරන රූඩ් 02 පර්චස් 10ක බිම් ප්‍රමාණයකින් යුතු ඉඩම.

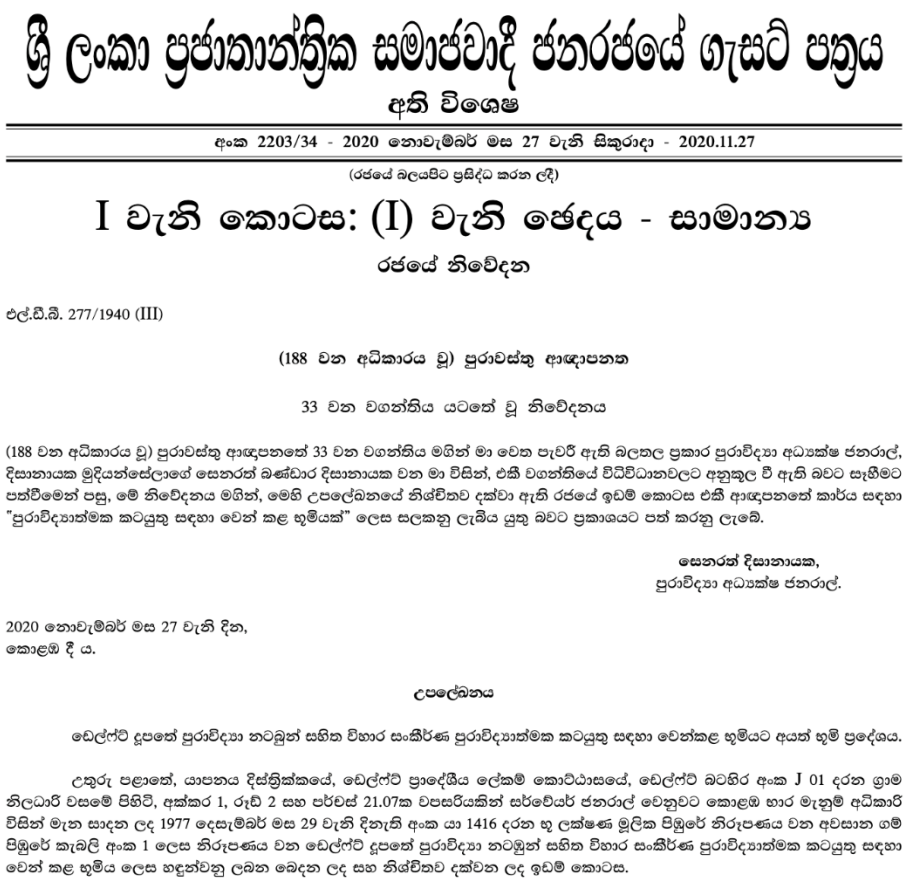
උතුරට	:	රුහුණ විශ්වවිද්‍යාලයේ ඉංජිනේරු පීඨ පරිශ්‍රය ;
නැගෙනහිරට	:	රුහුණ විශ්වවිද්‍යාලයේ ඉංජිනේරු පීඨ පරිශ්‍රය සහ පිඹුරුපත් අංක 5355හි කැබලි අංක B සහා C ;
දකුණට	:	පිඹුරුපත් අංක 5355හි කැබලි අංක J, K සහ L ;
බටහිරට	:	රුහුණ විශ්වවිද්‍යාලයේ ඉංජිනේරු පීඨ පරිශ්‍රය සහ කැබලි අංක 3.

10 - 642

Figure 12- Sample 1 gazette

Gazette No:	අංක 2195/35
Released Date:	2020 සැප්තැම්බර් මස 30 වැනි බදාදා
Issued By:	මහාචාර්ය ජී. එල්. පීරිස්, අධ්‍යාපන අමාත්‍ය
Issued Department:	අංක 18, වෝඩ් පෙදෙස, කොළඹ 07, අධ්‍යාපන අමාත්‍යාංශයේ දී ය.
Act(s):	ඉඩම් අත්කර ගැනීමේ පනතේ 49(අ) වගන්තිය
Section:	I වැනි කොටස I, වැනි ඡෙදය
Title	රජයේ නිවේදන
Sub Title	-
Notice:	රුහුණු විශ්වවිද්‍යාලයේ ඉංජිනේරු පීඨයේ විද්‍යුත් තොරතුරු ඉංජිනේරු අධ්‍යයනාංශයේ ඉදිකිරීමට යෝජිත ගොඩනැගිල්ල සඳහා පහත විස්තර වන ඉඩම ගැනීම අවශ්‍ය බව යටතේ මෙයින් නියම කරමි.

Sample 2:



ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය

අති විශෙෂ

අංක 2203/34 - 2020 නොවැම්බර් මස 27 වැනි සිකුරාදා - 2020.11.27

(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)

I වැනි කොටස: (I) වැනි ඡෙදය - සාමාන්‍ය

රජයේ නිවේදන

එල්.ඩී.බී. 277/1940 (III)

(188 වන අධිකාරය වූ) පුරාවස්තු ආඥාපනත

33 වන වගන්තිය යටතේ වූ නිවේදනය

(188 වන අධිකාරය වූ) පුරාවස්තු ආඥාපනතේ 33 වන වගන්තිය මගින් මා වෙත පැවරී ඇති බලතල ප්‍රකාර පුරාවිද්‍යා අධ්‍යක්ෂ ජනරාල්, දිසානායක මුදියන්සේලාගේ සෙනරත් බණ්ඩාර දිසානායක වන මා විසින්, එකී වගන්තියේ විධිවිධානවලට අනුකූල වී ඇති බවට සෑහීමට පත්වීමෙන් පසු, මේ නිවේදනය මගින්, මෙහි උපලේඛනයේ නිශ්චිතව දක්වා ඇති රජයේ ඉඩම් කොටස එකී ආඥාපනතේ කාර්ය සඳහා "පුරාවිද්‍යාත්මක කටයුතු සඳහා වෙන් කළ භූමියක්" ලෙස සලකනු ලැබිය යුතු බවට ප්‍රකාශයට පත් කරනු ලැබේ.

සෙනරත් දිසානායක,
පුරාවිද්‍යා අධ්‍යක්ෂ ජනරාල්.

2020 නොවැම්බර් මස 27 වැනි දින,
කොළඹ දී ය.

උපලේඛනය

ඩෙල්ෆ්ට් දූපතේ පුරාවිද්‍යා නටබුන් සහිත විහාර සංකීර්ණ පුරාවිද්‍යාත්මක කටයුතු සඳහා වෙන්කළ භූමියට අයත් භූමි ප්‍රදේශය.

උතුරු පළාතේ, යාපනය දිස්ත්‍රික්කයේ, ඩෙල්ෆ්ට් ප්‍රාදේශීය ලේකම් කොට්ඨාසයේ, ඩෙල්ෆ්ට් බටහිර අංක J 01 දරන ග්‍රාම නිලධාරී වසමේ පිහිටි, අක්කර 1, රූඩ් 2 සහ පර්චස් 21.07ක වපසරියකින් සර්වේයර් ජනරාල් වෙනුවට කොළඹ හාර මැනුම් අධිකාරී විසින් මැන සාදන ලද 1977 දෙසැම්බර් මස 29 වැනි දිනැති අංක යා 1416 දරන භූ ලක්ෂණ මූලික පිඹුරේ නිරූපණය වන අවසාන ගම් පිඹුරේ කැබලි අංක 1 ලෙස නිරූපණය වන ඩෙල්ෆ්ට් දූපතේ පුරාවිද්‍යා නටඹුන් සහිත විහාර සංකීර්ණ පුරාවිද්‍යාත්මක කටයුතු සඳහා වෙන් කළ භූමිය ලෙස හඳුන්වනු ලබන බෙදන ලද සහ නිශ්චිතව දක්වන ලද ඉඩම් කොටස.

Figure 13 - Sample 2 gazette

Gazette No:	අංක 2203/34
Released Date:	2020 නොවැම්බර් මස 27 වැනි සිකුරාදා
Issued By:	සෙනරත් දිසානායක, පුරාවිද්‍යා අධ්‍යක්ෂ ජනරාල්
Issued Department:	කොළඹ දී ය.
Act(s):	(188 වන අධිකාරය වූ) පුරාවස්තු ආඥාපනත
Section:	I වැනි කොටස I, වැනි ඡෙදය
Title	රජයේ නිවේදන
Sub Title	33 වන වගන්තිය යටතේ වූ නිවේදනය
Notice:	පැවරී ඇති බලතල ප්‍රකාර දිසානායක මුදියන්සේලාගේ සෙනරත් බණ්ඩාර දිසානායක විසින් එකී වගන්තියේ විධිවිධානවලට අනුකූල වී ඇති බවට සෑහීමට මේ නිවේදනය මෙහි නිශ්චිතව දක්වා ඇති රජයේ ඉඩම් කොටස එකී ආඥාපනතේ කායර් සඳහා – පුරාවිද්‍යාත්මක කටයුතු සඳහා වෙන් කළ භ+මියක්~ ලෙස සලකනු ලැබිය යුතු බවට ප්‍රකාශයට පත් කරනු ලැබේ.

Sample 3:

ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය

අති විශෙෂ

අංක 2228/11 - 2021 මැයි මස 17 වැනි සඳුදා - 2021.05.17

(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)

I වැනි කොටස: (I) වැනි ඡෙදය - සාමාන්‍ය

ජනාධිපතිතුමාණන් විසින් කරන ලද පත්කිරීම් ආදිය

2021 අංක 486/1

ජනා. කා. අංකය : පීඑස්/සීඑස්ඒ/00/1/8/2.

අතිගරු ජනාධිපතිතුමා විසින් ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ආණ්ඩුක්‍රම ව්‍යවස්ථාවේ 52(1) වැනි ව්‍යවස්ථාවේ බලතල ප්‍රකාර, ආණ්ඩුක්‍රම ව්‍යවස්ථාවේ 45(1) වැනි ව්‍යවස්ථාව යටතේ පිහිටුවන ලද ප්‍රජා පොලිස් සේවා රාජ්‍ය අමාත්‍යාංශයේ ලේකම් වශයෙන් එස්. ටී. කොඩිකාර මහතා 2021 මැයි මස 10 දිනැතිව, වහාම ක්‍රියාත්මක වන පරිදි පත්කරන ලද බව මෙයින් දැනුම් දෙනු ලැබේ.

අතිගරු ජනාධිපතිතුමාගේ නියමය පරිදි,

පී. බී. ජයසුන්දර,
ජනාධිපති ලේකම්.

2021 මැයි මස 17 වැනි දින,
කොළඹ 01,
ජනාධිපති කාර්යාලයේ දී ය.

Figure 14 - Sample 3 gazette

Gazette No:	අංක 2228/11
Released Date:	2021 මැයි මස 17 වැනි සඳුදා
Issued By:	පී. බී. ජයසුන්දර, ජනාධිපති ලේකම්
Issued Department:	කොළඹ 01, ජනාධිපති කායර්ාලයේ දී ය.
Act(s):	-
Section:	I වැනි කොටස I, වැනි ඡෙදය
Title	ජනාධිපතිතුමාණන් විසින් කරන ලද පත්කිරීම් ආදිය
Sub Title	-
Notice:	අතිගරු ජනාධිපතිතුමා විසින් ආණ්ඩුක්‍රම ව්‍යවස්ථාවේ බලතල යටතේ පිහිටුවන ලද ප්‍රජා පොලිස් සේවා රාජ්‍ය අමාත්‍යාංශයේ ලේකම් වශයෙන් එස්. ටී. කොඩිකාර මහතා 2021 මැයි මස 10 දිනැතිව වහාම ක්‍රියාත්මක වන පරිදි පත්කරන ලද බව මෙයින් දැනුම් දෙනු ලැබේ.

Sample 4:

ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය

අති විශෙෂ

අංක 2193/8 - 2020 සැප්තැම්බර් මස 15 වැනි අඟහරුවාදා - 2020.09.15

(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)

IV (ආ) වැනි කොටස - පළාත් පාලනය

පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනත යටතේ දැන්වීම්

(262 අධිකාරය) පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනතේ 66 ඇ (1)(ඇ) වගන්තිය යටතේ
බදුල්ල මහ නගර සභාවේ සභික ධුරයේ පුරප්පාඩුවක් පිරවීම

පක්ෂ සාමාජිකත්වය අහිමිවීම හේතු කොට ගෙන බදුල්ල මහ නගර සභාවේ සභික ධුරයක් පුරප්පාඩු වී ඇති හෙයින් ද ;

ඉහත කී ආඥාපනතේ 66 ඇ (1)(ඇ) වගන්තිය යටතේ බදුල්ල මහ නගර සභාවේ අංක 13 - කණුපැලැල්ල කොට්ඨාසයට තේරී පත්වූ බාලගේ ආතර් සිල්වා අයත් වන්නා වූ ශ්‍රී ලංකා පොදුජන පෙරමුණ නම් වන පිළිගත් දේශපාලන පක්ෂයේ ලේකම්ට දින තිහ (30) ක කාල පරිච්ඡේදය ඇතුළත පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනතේ (262 අධිකාරය) 9 වගන්තිය යටතේ නුසුදුස්සෙකු වී නොමැති අපේක්ෂකයෙකු, සභිකයෙකු වශයෙන් තෝරා පත් කර ගනු ලැබිය යුතු යැයි ප්‍රකාශ කරනු ලැබිය යුතු වූ ද, එම පක්ෂයට අයත් වන්නා වූ ද, එම බදුල්ල මහ නගර සභාවේ අංක 13 - කණුපැලැල්ල කොට්ඨාසයේ සභිකයා පිළිබඳව තීරණය කිරීමට මා විසින් නියම කරනු ලැබ ඇති හෙයින් ද ;

ඉහත කී පිළිගත් දේශපාලන පක්ෂයේ ලේකම් විසින් එම පුරප්පාඩුව පිරවීම සඳහා රත්නායක මුදියන්සේලාගේ නාලක නිශාන්ත විජයරත්න නම් කිරීමේ ස්වකීය තීරණය දන්වනු ලැබ ඇති හෙයින් ද ;

බදුල්ල මහ නගර සභාවේ තේරීම්භාර නිලධාරී, එච්. අයි. ආර්. හතුරුසිංහ වන මම, පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනතේ (262 අධිකාරය) 66 ඇ (1)(ඇ) වගන්තිය ප්‍රකාරව ක්‍රියා කරමින් රත්නායක මුදියන්සේලාගේ නාලක නිශාන්ත විජයරත්න, බදුල්ල මහ නගර සභාවේ අංක 13 - කණුපැලැල්ල කොට්ඨාසයේ සභිකයා වශයෙන් තෝරා පත්කර ගත් බව මෙයින් දැන් ප්‍රකාශ කරමි.

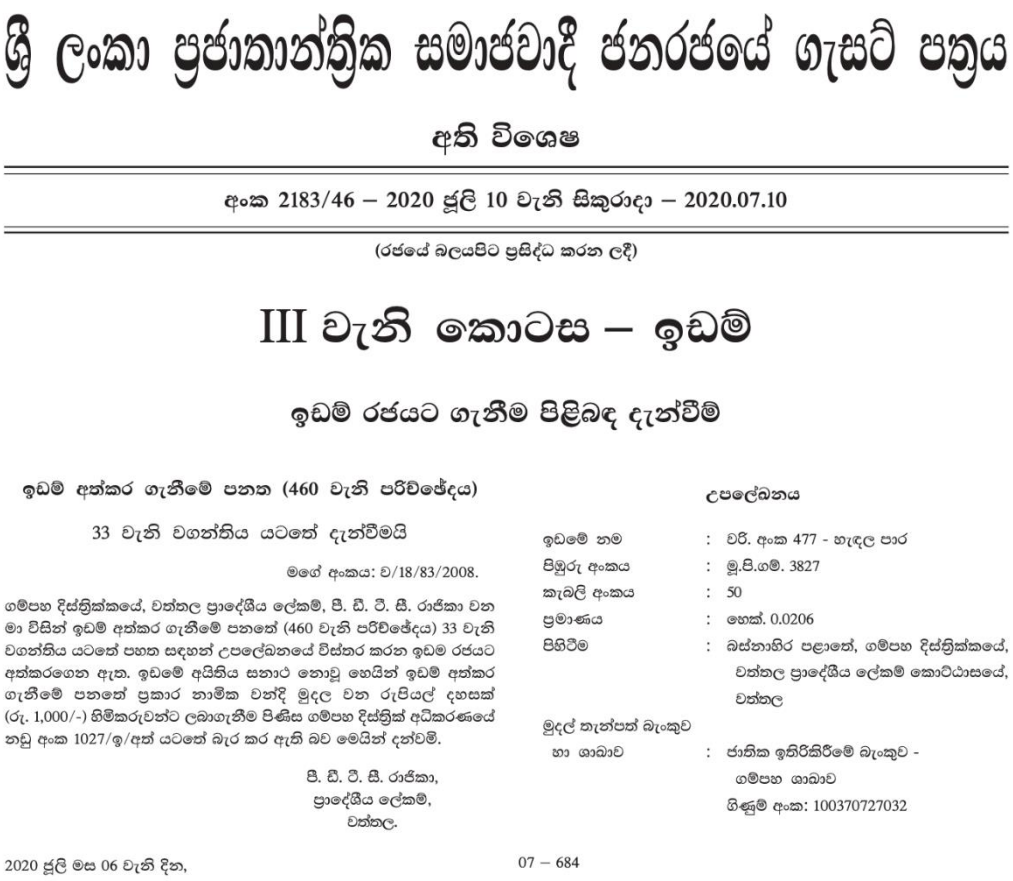
එච්. අයි. ආර්. හතුරුසිංහ,
තේරීම්භාර නිලධාරී,
බදුල්ල මහ නගර සභාව.

2020 සැප්තැම්බර් මස 14 වැනි දින,
දිස්ත්‍රික් මැතිවරණ කාර්යාලය,
බදුල්ල.

Figure 15 - Sample 4 gazette

Gazette No:	අංක 2193/8
Released Date:	2020 සැප්තැම්බර් මස 15 වැනි අඟහරුවාදා
Issued By:	එච්. අයි. ආර්. හතුරුසිංහ, තේරීම්භාර නිලධාරී
Issued Department:	දිස්ත්‍රික් මැතිවරණ කායර්ාලය, බදුල්ල.
Act(s):	(262 අධිකාරය) පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනතේ 66 ඇ (1)(ඇ) වගන්තිය
Section:	IV (ආ) වැනි කොටස , පළාත් පාලනය
Title	පළාත් පාලන ආයතන ඡන්ද විමසීම් ආඥාපනත යටතේ දැන්වීම්
Sub Title	බදුල්ල මහ නගර සභාවේ සභික ධුරයේ පුරප්පාඩුවක් පිරවීම
Notice:	පක්ෂ සාමාජිකත්වය අහිමිවීම හේතු කොට ගෙන බදුල්ල මහ නගර සභාවේ සභික ධුරයක් පුරප්පාඩු වී ඇති හෙයින්, බදුල්ල මහ නගර සභාවේ තේරීම්භාර නිලධාරී හතුරුසිංහ ප්‍රකාරව ක්‍රියා කරමින් රත්නායක මුදියන්සේලාගේ නාලක නිශාන්ත විජයරත්න මහ නගර සභාවේ අංක 13 කණුපැලැල්ල කොට්ඨාසයේ සභිකයා වශයෙන් තෝරා පත්කර ගත් බව මෙයින් දැන් ප්‍රකාශ කරමි.

Sample 5:



ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රය

අති විශෙෂ

අංක 2183/46 – 2020 ජූලි 10 වැනි සිකුරාදා – 2020.07.10

(රජයේ බලයපිට ප්‍රසිද්ධ කරන ලදී)

III වැනි කොටස – ඉඩම්

ඉඩම් රජයට ගැනීම පිළිබඳ දැන්වීම්

ඉඩම් අත්කර ගැනීමේ පනත (460 වැනි පරිච්ඡේදය)

33 වැනි වගන්තිය යටතේ දැන්වීමයි

මගේ අංකය: ව/18/83/2008.

ගම්පහ දිස්ත්‍රික්කයේ, වත්තල ප්‍රාදේශීය ලේකම්, පී. ඩී. ටී. සී. රාජිකා වන මා විසින් ඉඩම් අත්කර ගැනීමේ පනතේ (460 වැනි පරිච්ඡේදය) 33 වැනි වගන්තිය යටතේ පහත සඳහන් උපලේඛනයේ විස්තර කරන ඉඩම රජයට අත්කරගෙන ඇත. ඉඩමේ අයිතිය සනාථ නොවූ හෙයින් ඉඩම් අත්කර ගැනීමේ පනතේ ප්‍රකාර නාමික වන්දි මුදල් වන රුපියල් දහසක් (රු. 1,000/-) හිමිකරුවන්ට ලබාගැනීම පිණිස ගම්පහ දිස්ත්‍රික් අධිකරණයේ නඩු අංක 1027/ඉ/අත් යටතේ බැර කර ඇති බව මෙයින් දන්වමි.

පී. ඩී. ටී. සී. රාජිකා,
ප්‍රාදේශීය ලේකම්,
වත්තල.

2020 ජූලි මස 06 වැනි දින,

උපලේඛනය

ඉඩමේ නම	:	වරි. අංක 477 - හැඳල පාර
පිඹුරු අංකය	:	මූ.පි.ගම්. 3827
කැබලි අංකය	:	50
ප්‍රමාණය	:	හෙක්. 0.0206
පිහිටීම	:	බස්නාහිර පළාතේ, ගම්පහ දිස්ත්‍රික්කයේ, වත්තල ප්‍රාදේශීය ලේකම් කොට්ඨාසයේ, වත්තල
මුදල් තැන්පත් බැංකුව හා ශාඛාව	:	ජාතික ඉතිරිකිරීමේ බැංකුව - ගම්පහ ශාඛාව ගිණුම් අංක: 100370727032

07 – 684

Figure 16 - Sample 5 gazette

Gazette No:	අංක 2183/46
Released Date:	2020 ජූලි 10 වැනි සිකුරාදා
Issued By:	පී. ඩී. ටී. සී. රාජිකා, ප්‍රාදේශීය ලේකම්
Issued Department:	-
Act(s):	ඉඩම් අත්කර ගැනීමේ පනත (460 වැනි පරිච්ෙඡ්දය)
Section:	III වැනි කොටස , ඉඩම්
Title	ඉඩම් රජයට ගැනීම පිළිබඳ දැන්වීම්
Sub Title	33 වැනි වගන්තිය යටතේ දැන්වීමයි
Notice:	පහත සඳහන් විස්තර කරන ඉඩම රජයට ගෙන ඇත. ඉඩමේ අයිතිය සනාථ නොවූ හෙයින් පනතේ ප්‍රකාර නාමික වන්දි මුදල වන රුපියල් දහසක් හිමිකරුවන්ට ලබාගැනීම පිණිස ගම්පහ දිස්ත්‍රික් අධිකරණයේ නඩු අංක 1027/ඉ/අත් යටතේ බැර කර ඇති බව මෙයින් දන්වමි.

Appendix B

Tested F-Score Results for Manual Evaluated Gazettes

No	Sample File	No of Words	Abstractive Summarization	Extractive Summarization		Average F-Score
				Words	Keywords	
1	2189-48_S.pdf	73	0.91	0.39	0.61	0.64
2	2196-22_S.pdf	108	0.88	0.42	0.28	0.53
3	2181-24_S.pdf	91	0.82	0.48	0.34	0.55
4	2228-41_S.pdf	76	0.82	0.55	0.39	0.59
5	2193-31_S.pdf	58	0.85	0.41	0.55	0.60
6	2196-06_S.pdf	65	0.83	0.53	0.76	0.71
7	2196-43_S.pdf	98	0.81	0.18	0.43	0.47
8	2217-39_S.pdf	165	0.69	0.24	0.18	0.37
9	2188-29_S.pdf	52	0.88	0.57	0.8	0.75
10	2182-32_S.pdf	78	0.84	0.48	0.59	0.64
11	2045-06_S.pdf	119	0.86	0.61	0.39	0.62
12	2193-28_S.pdf	100	0.82	0.28	0.15	0.42
13	2211-25_S.pdf	169	0.86	0.2	0.24	0.43
14	2206-23_S.pdf	249	0.52	0.54	0.39	0.48
15	2211-60_S.pdf	53	0.67	0.66	0.39	0.57
16	2210-25_S.pdf	54	0.7	0.59	0.44	0.58
17	2210-60_S.pdf	162	0.86	0.21	0.18	0.42
18	2220-34_S.pdf	102	0.52	0.2	0.46	0.39
19	2179-04_S.pdf	165	0.63	0.51	0.1	0.41
20	2195-09_S.pdf	202	0.66	0.62	0.39	0.56
21	2203-30_S.pdf	206	0.83	0.33	0.16	0.44
22	2227-15_S.pdf	165	0.56	0.1	0.3	0.32
23	2216-41_S.pdf	205	0.52	0.17	0.29	0.33
24	2210-27_S.pdf	170	0.52	0.66	0.48	0.55
25	2210-62_S.pdf	86	0.66	0.53	0.35	0.51
26	2045-04_S.pdf	248	0.9	0.18	0.17	0.42
27	2223-04_S.pdf	239	0.73	0.39	0.3	0.47

28	2216-06_S.pdf	137	0.6	0.09	0.25	0.31
29	2230-11_S.pdf	184	0.53	0.44	0.4	0.46
30	2208-29_S.pdf	87	0.72	0.33	0.39	0.48
31	2183-48_S.pdf	85	0.56	0.51	0.13	0.40
32	2214-09_S.pdf	103	0.56	0.26	0.27	0.36
33	2195-36_S.pdf	69	0.69	0.24	0.14	0.36
34	2194-73_S.pdf	215	0.78	0.15	0.28	0.40
35	2197-04_S.pdf	100	0.66	0.56	0.1	0.44
36	2196-04_S.pdf	108	0.79	0.4	0.39	0.53
37	2181-02_S.pdf	72	0.62	0.47	0.18	0.42
38	2196-41_S.pdf	50	0.87	0.68	0.39	0.65
39	2192-52_S.pdf	225	0.79	0.33	0.35	0.49
40	2184-11_S.pdf	180	0.79	0.66	0.2	0.55
41	2180-26_S.pdf	193	0.9	0.4	0.19	0.50
42	2230-08_S.pdf	243	0.86	0.54	0.25	0.55
43	2231-08_S.pdf	235	0.52	0.37	0.17	0.35
44	2209-75_S.pdf	218	0.88	0.27	0.3	0.48
45	2228-06_S.pdf	216	0.63	0.58	0.36	0.52
46	2185-70_S.pdf	141	0.73	0.66	0.12	0.50
47	2229-06_S.pdf	217	0.83	0.52	0.19	0.51
48	2198-15_S.pdf	198	0.87	0.59	0.42	0.63
49	2223-20_S.pdf	65	0.6	0.22	0.47	0.43
50	2210-03_S.pdf	216	0.61	0.45	0.47	0.51