Journal of Science-FAS-SEUSL (2023) 04(01) 01-15
ISSN:2738-2184

Automated text summarization of Sinhala online articles

MAC Akmal Jahan and KKC Wijesekara

Department of Computer Science, Faculty of Applied Sciences,
South Eastern University of Sri Lanka.
Email: akmaljahan@seu.ac.lk

Abstract

Information retrieval is one of the major tasks in natural language processing applications. In digitalized world, there is a development of retrieval information from online platforms and there are abundant of information for a specific subject available in online. With the hustle and bustle, readers need to know whether the information is important according to their need within a very short time. Automated text summarization plays a key role in natural language processing applications. Many studies have been explored for summarizing different languages like English, Bengali, Hausa, Chinese, Hindi, etc. However, the local language like Sinhala is still in beginning stage. On the other hand, as a diverse country, there is a community and language diversity in Sri Lanka. Therefore, there are people who have less fluency in Sinhala as their mother-tongue is another local language like Tamil. Social media like Facebook provides platform for translation of content in a different language. However, other online platforms do not provide such translation process of the content. In such scenario, having a short summary of those articles would be an advantageous step for the readers who can easily understand the main idea of the content. Therefore, this work aims to generate an online platform that can provide a good summary for Sinhala language online articles. This research investigates extractive text summarization for Sinhala online articles using some state-of-the-art algorithms in NLP applications to select a best suitable method. This work comparatively analyses the performance of TF-IDF (Term Frequency-Inverse Document Frequency) and Text-Rank algorithms for Sinhala language. Performance of the algorithms is evaluated with human generated summary from online sources using ROUGE (Recall Oriented Understudy of Gisting Evaluation) where high ROUGE score (Measure the rate of n-gram overlapping of original text and automated summary) values represent the more accurate automated summary of the article. From the results, the TF-IDF algorithm comparatively performs better for Sinhala online article summarization with medium content size.

Keywords: Text summarization, Text-Rank, TF-IDF, Sinhala article

1. Introduction

Automated text summarization is a major domain in Natural Language Processing. The process of decreasing the content of a text by identifying significant meaningful information in a document known as text summarization. With the development of artificial intelligence and after the invention of machine translation, many research have been exploited for the text summarization

Corresponding Author: akmaljahan@seu.ac.lk

and it is becoming prevalent due to the fact that the manual text summarization is a time-consuming process and generally a laborious task. Text summarization plays a vital role and act as an intermediate stage in applications of various NLP related tasks. They can range from question answering, text classification, news summarization and headline generation etc. Since we are living in a fast-moving world in the present era, most of the people have tended to use online platforms to facilitate day to day activities and save their time. For instance, buying and reading printed newspapers is gradually decreasing in the younger generation who use online platforms and social media for the alternative sources. There is huge amount of information available in online platform in various formats like audio, video, text, images, etc. Among them, we mostly use online articles such as news articles, blogs, review, politics and featured articles to get day to day information. Since the hustle and bustle, people do not spend more time to read such articles, and majority of the mobile users skim and scroll and therefore skip some lengthy articles due to the time consumption. Sometime they waste their time to understand the context without knowing whether the context is relevant or irrelevant. The text headlines do not always express the real content of the text and users may fail in the middle of the reading. What if we have a summary of the text in such situation? Readers can easily understand the content of the article and decide if it is relevant to read further or skip. In this context, this research focuses on text summarization of a local language in online platform. Summarization of a text either be extractive or abstractive. Extractive summarization directly takes the sentences from the text according to the significance whereas abstractive summarization generates a new summary by using own/new words to raise the meaning of the content. There are several other international and local languages have been exploited the above techniques. For instance, many international languages like English, Arabic, Chinese, Bengali, Hindi, Husma etc. But the local language like Sinhala is in a opening state. Sinhala is a unique language for Sri Lanka that evolved from ancient languages, Pali and Sanskrit [2]. After passing many historical stages of language, in present we are using modern Sinhala language which enriches with a rich vocabulary and difficult grammatical rules. Because of that Sinhala is known as a complex language. As a diverse country, there is a community and language diversity in Sri Lanka. Therefore, there are people who have less fluency in Sinhala as their mother-tongue is another local language like Tamil. Social media like Facebook provides platform for translation of content in a different language. However, other online platforms do not provide such translation process of the content. In such scenario, having a short summary of those articles would be an advantageous step for the readers who can easily understand the main idea of the content.

Therefore, the aim of this work is to generate an online platform that can provide a good summary for Sinhala language online articles. To object the task in our research, we have investigated the extractive summarization algorithms for Sinhala local language to select a best suitable method. This remaining of the paper is organized as follows: Section II discusses related work and Section III describes the methods and technologies that aided for the summarization task of Sinhala language; Section IV discusses how the proposed algorithms performed for the summarization process and concluded in Section V.

2. Literature Review

From the beginning, written languages play an important role on documentation and pass on knowledge by securing information [1]. In today's world of digitalization, online methods have become a means of information retrieval for the purpose of saving time. Based on that fact, many studies have focused on the concept of text summarization [2]. The goal of the text summarization system is to produce a concise and coherent summary which would allow people to understand the concept of the input without reading the entire text [1]. In generally, text summarization can be classified into different types, depending on its aspects, namely, i). input type, ii). summary usage, iii) techniques iv) output type and v) approach [3]; and the most commonly used classification is the output type. Accordingly, there are two ways to summarize a corpus as, extractive and abstractive way.

Extractive summarization produces the summary, where information or sentences are extracted from the original document or given text file [4]. It is similar task to highlight the most relevant sentences from the original text or documents [2]. Main benefit of using extractive method is, it can choose the information or sentences that are significant and correct [2]. Also, no any modification included and it does not change the grammatical structure of the text [5][6]. First text summarization research is the extractive summarization method, introduced by Baxendale in 1958 by extracting important sentences using position of text. This work used some features as, term frequency and position in the text. The extractive method is famous method in the text summarization because of their simplicity [6].

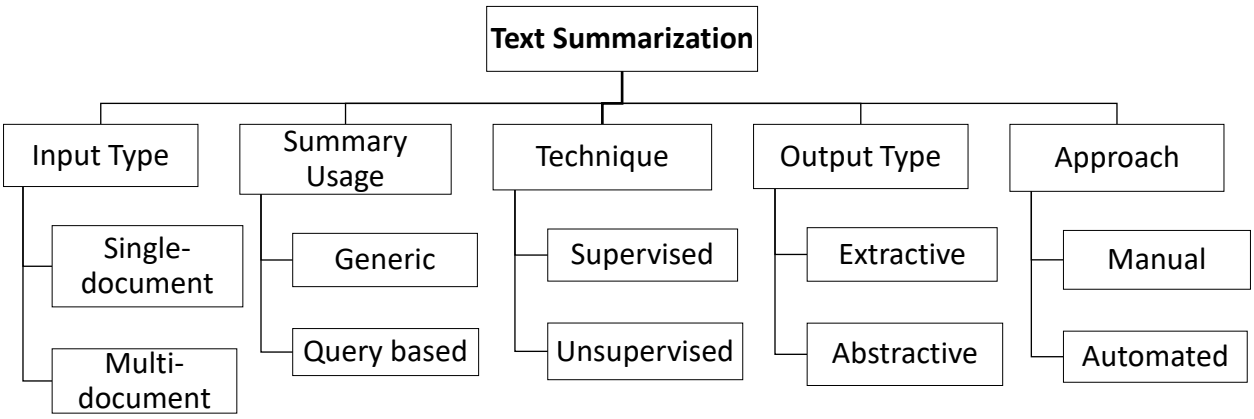

Figure1: Overall classification of text summarization

The main task in extractive summarization method is to select the important sentences using some special characteristics of the text content such as cue words, key words, title words, location, position words, sentence position, proper noun, sentence length, upper case, similarity and cohesion and term frequency [16]. For a better extractive summary, we need to rank the sentences

according to the importance of the words and then measure how one word relates to another word (similarity measure).

2.1 Text Summarization Algorithms

The existing studies for international languages used various text or sentence ranking algorithms such as Text Rank, TF-IDF, Seq2Seq and TFRSP, and they performed under different approaches like graph based, cluster-based, lexical chain-based approaches. Cosine similarity, word2vector, TF-IDF matrices are the numerical measures that have used to rank words or sentences and measure similarity of words. In our study, we have investigated the use of TF-IDF and Text Rank, algorithms for the extractive summarization in Sinhala language.

A. TF-IDF Algorithm

TF-IDF stands for Term Frequency-Inverse Document Frequency, a statical measure that evaluates the frequency of words. The main concept behind the algorithm is, extract the keywords in any content and rate the importance of that keywords based on how frequently they appear (scoring words). TF and IDF are the main components of the algorithm which act as matrices on finding relevant words by multiplying those two matrices [7], and the goal of this is to score the importance of a term in a corpus.

$$\text{TF} = \frac{No.of\ times\ word\ appears\ in\ the\ sentences}{No.of\ words\ in\ the\ sentences}$$

$$\text{IDF}=log\left(\frac{No.of\ sentences}{No.of\ sentences\ with\ the\ word}\right)$$

$$\text{TF-IDF}=\text{TF}*\text{IDF}$$

B. Text-Rank Algorithm

On the other hand, the Text Rank algorithm is an unsupervised method that uses for extractive summary of a text which intended from Google's page rank algorithm. Text Rank is a graph-based algorithm that can be used for key word extraction, automated text summarization and sentence ranking [2]. The concept behind that technique is words with higher frequency have wider relevance and are considered the most important phrases. Hence, the highly frequently words in sentences are more important than the others. The algorithm assigns grades to each sentence in the text based on this. The final summary only contains the top-ranked sentences of the articles.

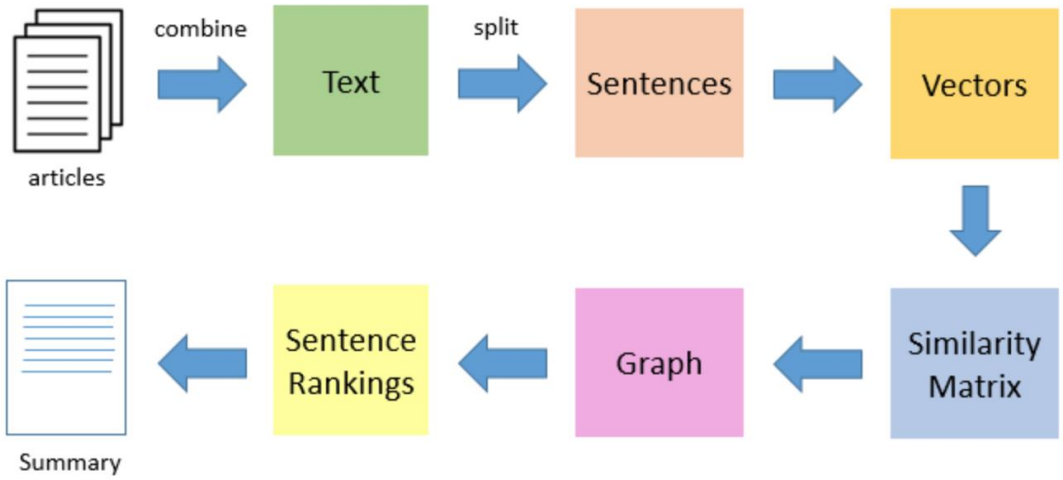

Figure **Error! No text of specified style in document.**: Overall Process of Text Rank Algorithm

To perform a text rank algorithm, there are few processes to be carried out as given below [8]:

i. Break the text corpus into individual sentences.
ii. Find vector representation (word embeddings) for each and every sentence.
iii. Generate a matrix that contains comparisons between the sentence vectors.
iv. Calculate sentence rank.
v. Transform the similarity matrix into a graph with sentences as nodes and similarity ratings as edges.
vi. Specific number of the top-ranked sentences will be composed in the final summary.

C. Sequence2Sequence Algorithm

Seq2Seq is a neural network based summarization algorithm that mostly involves for abstractive summarization but can use for the extractive summarization as well [17]. It is an encoder-decoder based model where the source corpus is encoded into context vector that preserves the text information. Target words are generated by the decoder for each time step according to the context vector [18]. Google has initiated this model for the applications of google translate, image capturing, online chat bots, text summarization etc. As the term sequence to sequence, the algorithm helps to generate new phrases by retaining the meaning of the content [2].

D. *SummerRUNNer Algorithm*

SummerRUNNner is a RNN based seq2seq model for extractive summarization of document [1]. It is a simple classifier without decoder, outperforming of matching the model by finding a set of sentences which the highest ROUGE with respect to the manual summary.

2.1 NLP for Sinhala Language

This study was primarily based on the Sinhala language articles. Sinhala is one of a richest language in the world [11]. It is one the native language of Sri Lanka. Due to various historical facts modern Sinhala language have used words from other languages such as English, Portuguese and Dutch[12].

Different NLP resources have discussed for Sinhala language by some researches. Sinhala word embedding has performed and evaluated with word2vec, fast text and glove vectors [13]. List of Sinhala stop words has introduced as result of derivation of those words from large corpora [11][14]. Many studies have been done for Sinhala NLP optical character recognition and also Sinhala-to-English translation. However, the facts are, Sinhala NLP research studies are in very much lagging behind compared to advancement in other international languages like English, Chinese, French, German.

3. Methodology

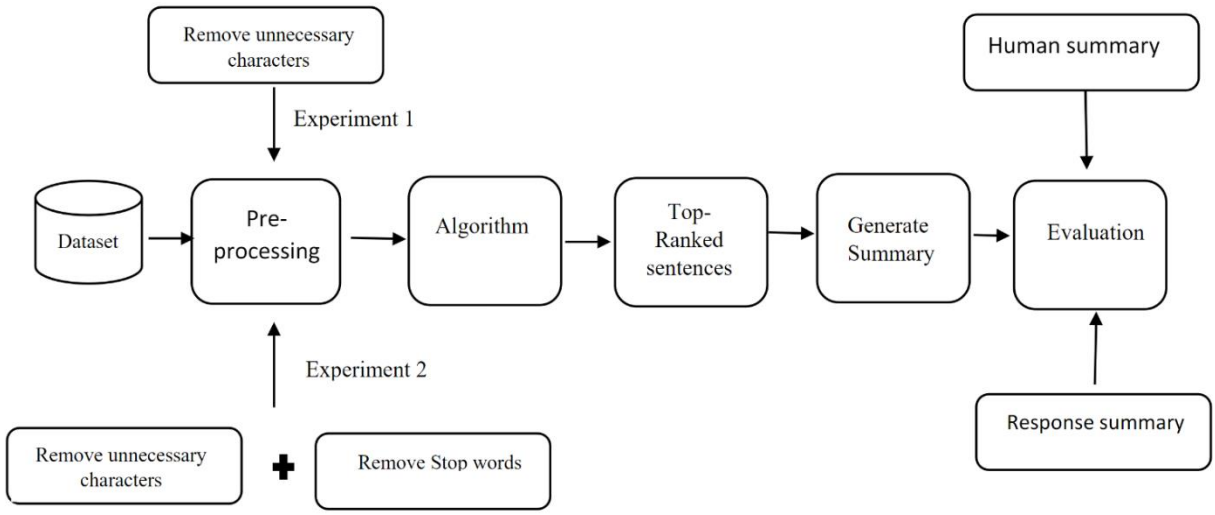

Figure 3: Overall Flow of the Text summarization Process

3.1 Evaluation Methods

A. ROUGE- Recall Oriented Understudy of Gisting Evaluation

Since there are many approaches and algorithms exploited in the text summarization , which needs an evaluation method to select the most accurate and efficient approach to compare and evaluate their performance [1]. ROUGE is a matric, used in machine translation and text summarization and measure n-gram overlapping between reference and automated summary which is generated

by human [2]. The Score is performed as ROUGE-1, ROUGE-2 and ROUGE-L. ROUGE-1 and ROUGE-2 are unigram and bigram overlap, which intended to measure the informativeness, while ROUGE-L denotes Longest common sequence which captures fluency to some extend[1]. These different types of ROUGE score values consist with another three measurements such as Precision, Recall and F-measure [9].

$$\text{Recall} = \frac{(Number\ of\ Words\ Overlapped)}{(Total\ words\ in\ human\ reference\ summary)}$$

$$\text{Precision} = \frac{(\text{Number of words overlapped})}{(\text{Total words in system summary})}$$

$$\text{F-measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Recall refers how much words of template summary is resembled to the automated summary whereas Precision denotes how much reference summary words are relevant to the automated summary. On the other hand, F-measure describes the whole story or harmonic mean of precision and Recall.

B. Word Embeddings-Text Vectorization

A computer cannot understand the input text, letters or any other characters and it can understand the numerical values. Therefore, human have been made commands to convert those characters into understandable format (numerical format). Word vectorization is the process that convert words into numbers where word or phrases from vocabulary are mapped to a corresponding vector of real numbers which used to find word predictions, word similarities or semantics. After words are converted to the vector, there is a need to use some methods to identify the similarity of words like Euclidean distance, Jaccard distance, cosine similarity, word mover's distance method. Cosine similarity is one of the words embedding methods where it measures the similarity between two non-zero vectors of an inner product space. Mathematically, it measures the Cosine of the angle between vector projection in a multi-dimensional space. Since the sentences of a text will be representing as the bunch of vectors, we can use it to find the similarity among the sentences [10]. In this work, we have used Cosine similarity method to identify the similar words for Text rank algorithm.

$$\text{Similarity} = \cos\theta = \frac{A.B}{||A||\,||B||}$$

3.2 Dataset Preparation

Finding available online data in Sinhala language for text summarization has proven to be quite challenging because of the automated text summarization for Sinhala language still in beginning

state. Therefore, no any standard data set available for Sinhala language. So, that our data set was limited to 300.

We used the Sinhala news articles for our research that collected from social media platforms and other online platforms that available Sinhala news articles. Statically, we collected 300 Sinhala news articles from Hiru News, Lankadeepa, Mawbima, Lakbima, Divayina and parliament news websites.

Then the collected articles separated into three categories as short (sentences 1-10), medium (sentences 10-30) and long articles (sentences more than 30) as equal amount from each category (100) for the purpose of checking the performance of algorithms according to the content size.

Human generated manual summaries are prepared for each article for evaluation task., by a group of university students who studying Sinhala as a subject and supervised by professional Sinhala teacher.

3.3 Data Pre-processing

Generally, text is organized in unstructured forms and consists of noise in different forms such as emotions, punctuation, text in a different case and therefore, it is too complex to deal with human language. This needs text preprocessing to clean the text corpus and make it ready to feed data to the model.

In our experiment, input dataset pre-processed with various stages. First, whole corpus is tokenized into words. From that collection of words, we have removed unnecessary characters (currency numbers, punctuation marks etc). Eliminating stop words is one of the significant pre-processing phases in many NLP applications. In that sense, articles and pronouns are classified as stop words. In this case, we have simply removed commonly occurring words in the corpus. Stemming performed using Porter-stemmer algorithm for reducing inflection in words to their root form while Lemmatization proceeds to change the words into their dictionary form [1][14].

4. Experiments

Part of this research mainly focused on the impact of stop word removal in Sinhala language corpus and investigate how much it affects the text summarization process. Therefore, we have performed two main experiments: i) handling text with stop words removal and ii) without stop word removal for the purpose of investigating how stop word removal in Sinhala language affects to yield better summarization [15].

A. Experiment 1- Stop words removal

Stop words are not very discriminative and have least significance in information related NLP applications, On the other hand, in some NLP applications they might have little impact. Therefore, the impact of these words is measured by removing them within the corpus for Sinhala language.

B. Experiment 2- without stop word removal

Sinhala is a unique language for Sri Lanka and it is difficult and different from other languages due to the involvement of several other languages and enriched with complex grammatical rules. Therefore, ranking sentences with removing words may affect to the final output summary. Similar to other languages, Sinhala language also shows discourse integration. When removing the stop words from the content, then the content may become meaningless text. According to the definition that stop words have no significance, but they directly affect the Sinhala language to improve the relationship between words in the construction of a meaningful sentence. Therefore, the corpus is treated with all existing words in this experiment.

Summary generation of the articles is implemented using python 3.10 and NLTK libraries. Initial corpus and the output text after using TF-IDF and Text rank algorithms are shown in Figures 4, 5 and 6 respectively.

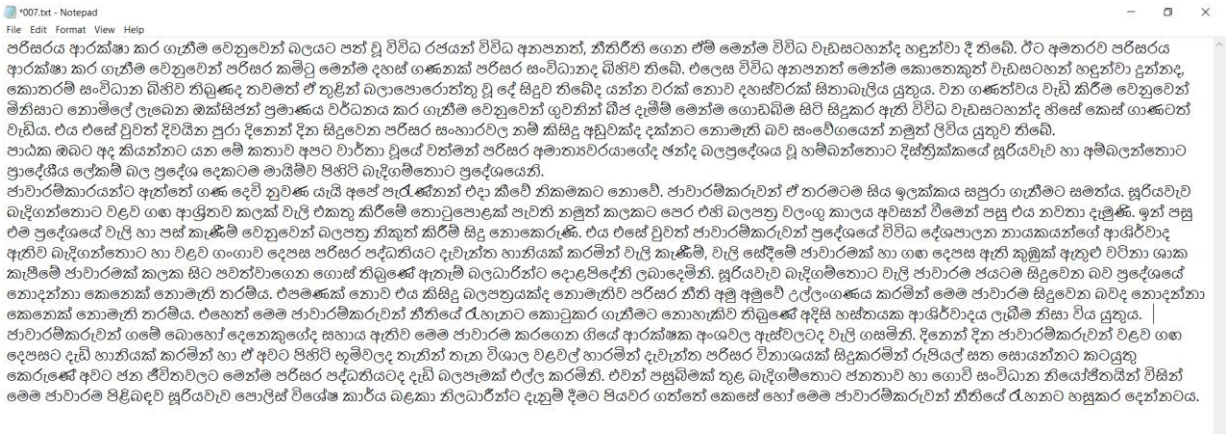

Figure 4: Input text from online news article

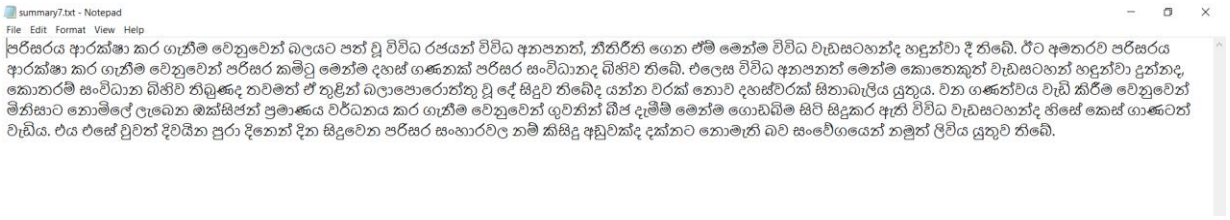

Figure 6: Output summary of Text Rank Algorithm

Figure 6: Output summary of Text Rank Algorithm

Recall, Precision, F-measure values of ROUGE-1, ROUGE-2 and ROUGE-L are used to determine, how much pertinent information contain from the initial text. Recall describes how much words of candidate summary are extracted and Precision says how much candidate summary words are relevant. But the final decision about the evaluation is made according to the F-measure value. Average of F-measure values have used for the analysis the results.

C. Experiment 3:

The experiment is carried out using content size of the article where short medium and long size articles are processed with both algorithms.

5. Results and Discussion

i. Performance Comparison of Experiments 1 and 2

From the experiments 1 and 2, the results are compared to check how stop words removal effect the automated summary generation of Sinhala language using TF-IDF and Text rank algorithms and the results are given in Table 1.

Table 1: N-gram comparison for stop word removal for both algorithms.

Experiment	Algorithm	Unigram	Bigram	LCS
1	TF-IDF	0.482156	0.373519	0.478262
	Text Rank	0.185981	0.109308	0.181606
2	TF-IDF	0.441803	0.310891	0.439054
	Text Rank	0.200595	0.093626	0.193969

Based on the results in Table 1, using TF-IDF algorithm in experiment 1 shows high f-measure value for unigram, bigram and LCS overlapping. The algorithm gives the output according to the frequency of the words, and stop words are also counted as important words. Therefore, they come to the final output and involve for the overlapping.

The text rank algorithm in experiment 2 shows high f-measure value for unigram. The algorithm performs as ranking the sentences. It extracts the full sentences when produce the summary. Therefore, overlapping automated summary with reference summary is minimal.

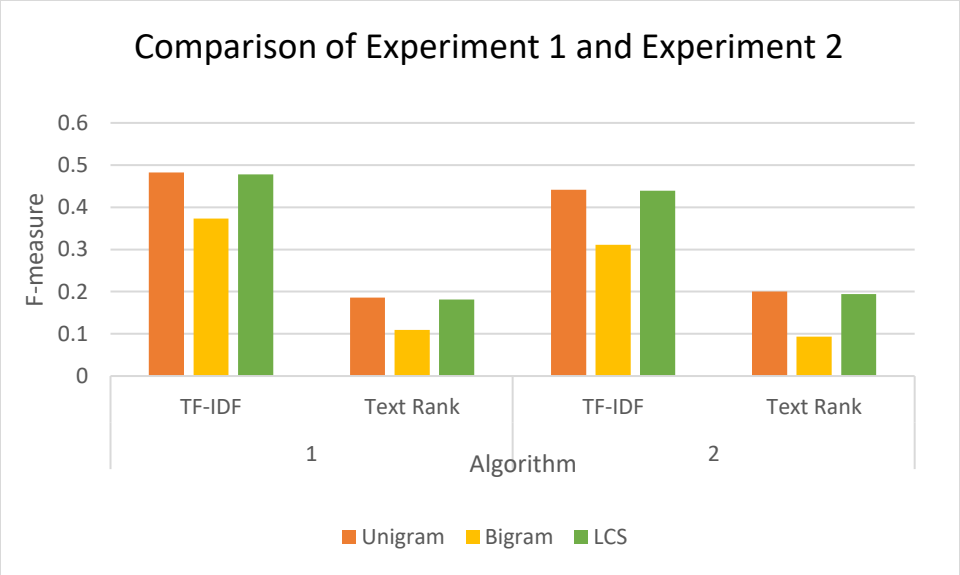

Figure 5: Comparison of Experiment 1 and Experiment 2

ii. Performance of Algorithms using Text Content Size

The experimental results given in Table 2 show that F-measure score for medium content size articles in both experiments using TF-IDF and Text Rank algorithms is higher for the extractive summarization of Sinhala articles. Therefore, Medium size content showed the best performance for both algorithms as shown in figure 6.

Table 2: Results of experiments 1 and 2 according to the content size.

Experiment	Algorithm	Content Size	Recall	Precision	F-Measure
1	TF-IDF	Short	0.501523	0.375095	0.399575
		Medium	**0.86551**	**0.538095**	**0.641731**
		Long	0.652178	0.386197	0.29263
	Text-Rank	Short	0.101549	0.011666	0.019443
		Medium	0.358735	**0.219909**	**0.267295**
		Long	**0.404263**	0.133495	0.190157
2	TF-IDF	Short	0.412311	0.354298	0.35621
		Medium	**0.706195**	**0.848926**	**0.557116**
		Long	0.628535	0.196013	0.278423
	Text-Rank	Short	0.184607	0.152951	0.151254
		Medium	**0.202561**	**0.151429**	**0.170487**
		Long	0.344779	0.120665	0.166448

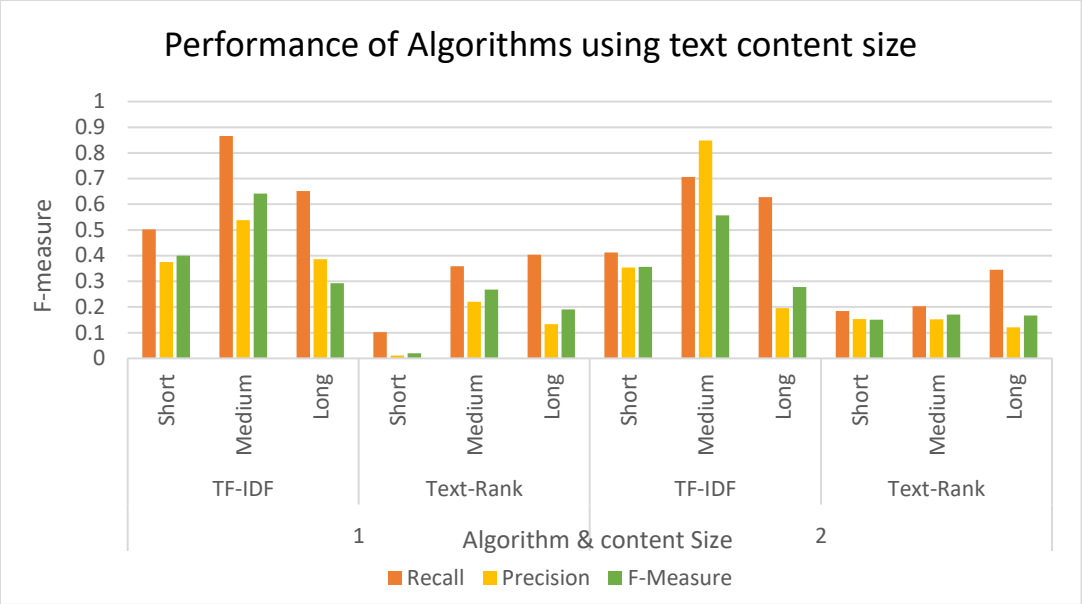

Figure 6: Performance of algorithms using text content size

iii. Overall Comparison of Algorithms

For the comparison of overall performance of TF-IDF and Text-Rank algorithms analysis using the average values of each n-gram's F-measure scores as shown in Table 1. According to the value variation shown in figures 7 and 8, TF-IDF algorithm (blue line) shows the high F-measure score than Text Rank algorithm (Red line) in both experiments.

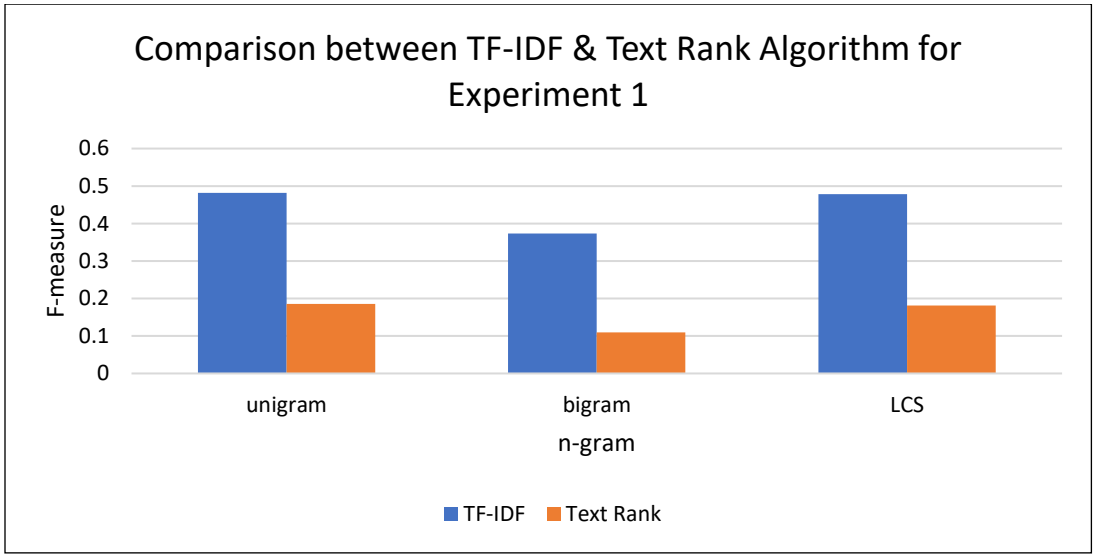

Figure 1: Comparison between TF-IDF & Text Rank Algorithms in experiment 1

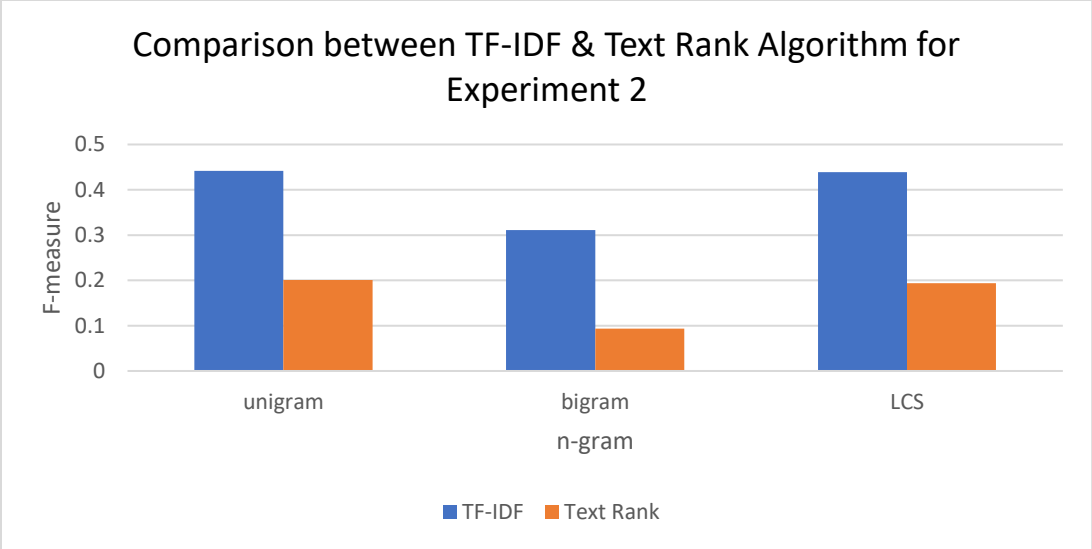

Figure 2: Comparison between TF-IDF and Text Rank algorithms in experiment 2

6. Conclusion

By the overall comparison of the algorithms in both experiments, the TF-IDF algorithm comparatively performs better for Sinhala online article summarization with medium content size. In the investigation of stop words removal in the text corpus, the process slightly influences the performance of the both algorithms.

The study has used a set of manual summaries of each online article corpus for the evaluation task. Since producing a summary is a matter of individual intelligence of a person, some manual summaries are in abstractive format. Therefore, in some cases the ROUGE scores show a very low value for n-gram overlapping. This research incurs further investigation of standardize manual summaries in future.

References

1. O. Klymenko, D. Braun, and F. Matthes, “Automatic text summarization: A state-of-the-art review,” in *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, 2020, vol. 1, pp. 648–655..
2. S. M. Meena, M. P. Ramkumar, R. E. Asmitha, and G. S. Emil Selvan, “Text Summarization Using Text Frequency Ranking Sentence Prediction,” Sep. 2020. doi: 10.1109/ICCCSP49186.2020.9315203.
3. J. R. Thomas, S. K. Bharti, and K. S. Babu, “Automatic keyword extraction for text

summarization in e-newspapers," *ACM Int. Conf. Proceeding Ser.*, vol. 25-26-Augu, no. April, 2016, doi: 10.1145/2980258.2980442.

4. S. Abujar, A. Kaisar Mohammad Masum, M. Mohibullah, and S. Akhter Hossain, "An Approach for Bengali Text Summarization using Word2Vector."

5. A. Rozaimee, M. Bashir, W. Malini, and W. Isa, "Automatic Hausa LanguageText Summarization Based on Feature Extraction using Naïve Bayes Model," *World Appl. Sci. J.*, vol. 35, no. 9, pp. 2074–2080, 2017, doi: 10.5829/idosi.wasj.2017.2074.2080.

6. A. Aries and W. K. Hidouci, "Automatic text summarization : What has been done and what has to be done," no. February, pp. 1–36, 2017.

7. K. A. M. P. Rathnasena, K. M. S. J. Kumarasinghe, D. T. P. Paranavitharana, and D. V. A. U. Dayarathne, "Summarization based approach for Old Sinhala Text Archival Search and Preservation," pp. 182–188, 2018.

8. I. F. of E. Christ University (Bangalore and Institute of Electrical and Electronics Engineers, 2019 International Conference on Data Science and Communication (IconDSC) : Faculty of Engineering, CHRIST (Deemed to be University), Bangalore, 2019-03-01 to 2019-03-02.

9. E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," *Lang. Resour. Eval.*, vol. 52, no. 1, pp. 101–148, 2018, doi: 10.1007/s10579-017-9399-2.

10. A. R. Deshpande and L. L.M.R.J, "Text summarization using clustering technique and SVM technique," *Int. J. Appl. Eng. Res.*, vol. 10, no. 10, pp. 25511–25519, 2015.

11. A Sinhala and Tamil Extension to Generic Environment for Context-aware Correction, Institute of Electrical and Electronics Engineers, *2019 National Information Technology Conference (NITC).*

12. N. De Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," pp. 1–24.

13. D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, "Word embedding evaluation for Sinhala," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 1874–1881, 2020.

14. Y. Wijeratne and N. De Silva, "Sinhala Language Corpora and Stopwords from a Decade of Sri".

15. "What Is Automatic Text Summarization? | Frase." https://www.frase.io/blog/what-is-automatic-text-summarization/ (accessed Aug. 29, 2022).

16. A. Sahoo, D. Kumar Nayak, and M. Tech Student, "Review Paper on Extractive Text Summarization," *Int. J. Eng. Res. Comput. Sci. Eng.*, vol. 5, no. 4, pp. 2394–2320, 2018.

17. W. Xiao and G. Carenini, “Extractive summarization of long documents by combining global and local context,” *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3011–3021, 2019, doi: 10.18653/v1/d19-1298.

18. Y. Zhang, D. Li, Y. Wang, Y. Fang, and W. Xiao, “Abstract text summarization with a convolutional seq2seq model,” *Appl. Sci.*, vol. 9, no. 8, 2019, doi:

19. A. Cohan and N. Goharian, “Revisiting summarization evaluation for scientific articles,” *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 806–813, 2016.

20. S. Banerjee and A. Lavie, “METEOR: An automatic metric for mt evaluation with improved correlation with human judgments,” *Intrinsic Extrinsic Eval. Meas. Mach. Transl. and/or Summ. Proc. Work. ACL 2005*, no. May, pp. 65–72, 2005.