ISSN 2347 - 3983
Volume 10. No.6, June 2022
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter071062022.pdf
https://doi.org/10.30534/ijeter/2022/071062022

Extractive Text Summarization for Wolaita

Aklilu Elias Kurika[1], Tigist Simon Sundado[2], Michael Melesse[3], Workineh Menna Eligo[4]
[1] Department of Information Technology, School of Informatics, Wolaita Sodo University, Ethiopia, akliluelias123@gmail.com
[2]Department of Information Technology, School of Informatics, Wolaita Sodo University, Ethiopia, simontig123@gmail.com
[3]Addis Ababa University, Ethiopia, Michael.melese@aau.edu.et,
[4]Department of Statistics, College of Natural and Computational Science, Wolaita Sodo University, Ethiopia, worke2003@gmail.com

Received Date : May 6 , 2022 Accepted Date : May 30, 2022 Published Date : June 07, 2022

ABSTRACT
Text summarization is the mechanism of summarizing a huge document comprising vast amount of information which is difficult to overcome and understand its message easily in any written documents for whatever languages without losing its entire message. A short and precise document which conveys intended information for the user in demand is expected in this information age. In addition to that, summarizing a document with vast amount of information is very difficult and time consuming specially for less resourced and technologically unfavored languages. Therefore in this study, the researcher proposed to address such problems for Wolaita by using graph based extractive text summarization approach. To attain the goal of this study the researcher prepared 92 documents for the study, explored extractive text summarization with graph-based approach to address the problems, performed text preprocessing tasks and finally developed text summarization model by using TextRank algorithms. The researcher used 92 documents, performed 92 various experiments, on documents and experimental results and findings were discussed in detail. To evaluate the model performance, three different expert summaries were collected for documents and computed system generated summaries with ROUGE evaluation metric. The researcher justified it with ROUGE evaluation metrics by comparing the system summaries with the expert summaries. The result obtained from the experiment shows promising result in summarization of Wolaita text. Finally, the experimental result of a 61.16% recall, 60.69% precision and 60.46% f-measures were obtained.

Key words: Text Summarization, Wolaita Language, Extractive Summarization, Graph based Approach, TextRank Algorithm, ROUGE Evaluation metric.

1. INTRODUCTION
Natural language is a communication means in which humans use to share ideas one another and natural language processing (NLP) is the automatic or semi-automatic process of human language [1]. It is also considered as the linguistic data analysis, generally in the form of textual data such as documents or publications, using computational methods. The goal of Natural language processing is to build a depiction of the text that adds structure to the unstructured natural language, by taking the linguistics benefits or insights [2]. This structure might be syntactic in nature capturing the more semantic capturing or the meaning conveyed by the text grammatical relationships among constituents of the text using the state of the art language technology [3]. Language Technologies are dedicated for dealing with human language which is the most complex information medium in our world and are also often subsumed under the term Human Language Technology [4]. As a result of advancing technology, different kind of digital resource are being generated in different languages in our daily usage at alarming rate even without noticing the existence of NLP application. The availability of these resources requires different NLP tasks. These NLP applications include, Text Summarization, Information retrieval (IR), Information Extraction (IE), Text classification, Machine translation (MT), Automatic Speech Recognition (ASR) among others [5]. Among these NLP applications, Text Summarization refers to the task of presenting information in a concise manner focusing on the most important parts of the data whilst preserving the meaning. The main idea of summarization is to find a subset of data which contains the “information” of the entire set. In today’s world, data generation and consumption are exploding at an exponential rate. Due to this, text summarization has become the necessity of many applications such as search engine, business analysis, market review etc. Automatic Document summarization involves producing a summary of the given text document without any human help. Text summarization in general categorized in to two classes according the number of documents given for the summarizer as a single document summarization and multi document summarization. If the given or input document is single, it is said to be single document summarization and if a given input text for summarization of more related documents, then it is said to be multi document summarization [2]. Ii is also broadly classified as extractive and abstractive Summarization

2. STATEMENT OF PROBLEMS
Statement of problems currently in the world are, the written documents getting more and more from day today

this resulted information overloading which is today's most relevant problem [3]. This results in wasting of time and budget to get the insight about the document to read. In order to get short, precise and condensed information with full meaning and coherence, information reduction mechanism is required. Therefore a mechanism to reduce documents without losing its intended information is expected. This in turn led us to develop particular text summarization techniques. A lot of researches and studies have been made for summarizing written texts for some of Ethiopian languages such as Amharic, Afan-Oromo and Tigray languages. Text summarization for Wolaita language has not been developed. Related works in text summarization for Ethiopian languages are poor by the performance and accuracy. Another problem is also reading unsummarized or the total document is time consuming and boring task and it is also difficult to understand the message of the document accurately. In reality, single document summarization is a challenging task; this is due to it requires detailed understanding of the original document and the summarized documents are still far from human summarization performance [4]. The method used by the authors [5] is a novel extractive summarization approach intended for long documents by integrating the local context of each topic along with the global context of the entire document. Their achievement is good while compared with the previous works but there is a redundancy problem. To deal with redundancy the authors recommended the future scholarly researchers to explore artificial neural networks and convolutional neural network methods which would be beneficial to mix explicit features, like salience and sentence position, into the conducted studies' neural approach. Another study which is entitled as Automatic text summarizer for Tigrigna language by [6] in 2017 used the ranking and extraction mechanism from the original document. But it has its own limitations by the accuracy of the developed model. It performs by looking the frequency of words from the sentence and by identifying title words of the document from the original document respectively. Finally they revealed that the title identification method is better that the frequent word extraction with registered recall, precision and F-Score values were 0.46(46%), 0.50(50%) and 0.48(48%) respectively actually the performance is less as we can see from the result. In addition to that the authors face difficulties to get documents for the study and their findings is also not good as it can be recognized from the result. These are limitations and gaps of previous works for text summarizations. So, by using gaps and limitations of previous works in addition to statement of problems, the researchers are planned to develop text summarization for Wolaita language by using particular extractive text summarization for Wolaita by using graph based approach. Thus the main aim of the study is to design an automatic text summarization for the Wolaita language. To attain this goal, activites like review related works to understand the state-of-the-art in natural language, text summarization and Graph based approach, collect and prepare representative documents for Wolaita text summarizer, design the architecture of text summarization, extract a feature for the development of automatic text summarizer, develop extractive text summarization model for Wolaita language, evaluate the performance of Wolaita text summarizer model and the finding or final results were reported and further research directions were recommend for the future interested scholars.

3.REVIEW OF RELATED WORKS

Table 1: Review of Related Works

No	References	Approaches	Findings
1	[7]	A novel approach by considering semantic relationships among the sentences of the document and used a novel approach by considering semantic relationships among the sentences of the document again and also Maximum Marginal Relevance (MMR) to re-rank sentences.	The author revealed optimization based, clustering based, word embedding based methods, information and item set based techniques for generating extractive summary fail to determine the redundant information generated from the summarized sentences. Then the author proposed novel based approach; therefore redundancy in the summarized sentence was employed by cosine similarity measurement based approach. Then performance of their work was improved compared to the existing works.
2	[8]	Optimization based, clustering based, word embedding based, information and item set based, methods	The authors identified that "topic of text and linguistic properties are not determinant for extractive text summarizations".
3	[9]	Statistical approach to extractive summary for sport data	The authors' achievement is 73%. These authors recommended the future researchers to use neural network approach to develop extractive summary with improved accuracy.
4	[10] & [11]	Cue method to calculate relevance of sentence based on the presence or absence of cue words in the dictionary and title method which computes the	The author reveals that extractive summarization approach is give best results every times while compared to abstractive approaches this is because abstractive approaches

		weight of sentence.	cope with problems such as semantic representation, inference and natural language generation which are relatively harder that data driven approaches like sentences extraction.
5	[12]	The author used neural model.	Their findings show that the final model achieved strong performance than the previous systems and base lines.
6	[13]	Only machine learning approaches like graph representation such as LexRank or neural network based Netsum.	The most accurate sentences were marked by FRQ algorithm than JAC and DEN methods; The frequent patters in the document contribute to the quality of text summarizer (TS); the developed model performance is also similar to the previous model, precision is better that the previous works but recall is poor while compared to the previous works
7	[14]	Machine Learning Approaches	This author observed that not all frequently occurring words are relevant; the author also suggested the use of stopword filter. Graph mining tool is recommended to be used than DGR Miner to search more specific types of patters.
8	[15]	Machine Learning	Extractive text summarization is popular and its result is better than that of abstractive summarization approach

Table 1 shows the summarized form of review of related works which have been used as references in this article. It shows similar works conducted for text summations for various languages, approaches used for the studies, and results or final findings for respective works.

Table 2 : Type, source, amount of data & tools used for related works in the above table

Ref. No	Type and Source of Data	Amount of Data	Tools used for study
[6]	Tigraigna news article collected from the sources of Aiga forum and dmtse Woyane tigray web sites	60 News articles	Python Language
[16]	English corpus is gained from DUC in 2002 which means Document Understanding Conference.	50 reference sets (5-15) documents	ASP.Net and C# Language
[17]	Amharic news article; 47 from Melese (the previous researcher) and 13 new articles from Ethiopian reporter website.	60 News articles	Java Language
[18]	Amharic news articles from Ethiopian Reported, Amharic version.	50 News articles	Java Language
[19]	Tigrigna news from Aiga forum, Dimtse Weyane Tigray and Tigray television websites in .txt format.	120 news article	Python language
[20]	350 Cue phrases are collected from primary and secondary sources and translated to 729 Afaan Oromo cue phrases	350 English & 729 Oromo phrases	C# language is used

Table 2 shows the type of data, source of data, amount of data & tools used for related works represented above in table 1.

4. METHODOLOGIES OF THE STUDY

4.1 Literature Review: in order to identify what has been done and what is expected to be addressed in the areas.

4.2 Data Collection: text datasets were collected from Wolaita text books of primary school.

4.3 Tools: python programming language by Jupytor Notebook (Anaconda) editor

4.4 Techniques: Extractive Summarization with graph based approach.

4.5 Algorithm: TextRank Algorithm.

4.6 Performance Evaluation: ROUGE evaluation metric. The accuracy of model was measured by calculating precision, recall and f-measure. The study methodology used for this research is extractive text summarization approach which summarizes texts by selecting the most important sentences from the given document for summarization [16]. The rationale to select this approach is that various scholars and researchers recommend using

this approach to develop text summarization model with better accuracy. Abundant of the automatic text summarization researches focus on the extractive summary because abstractive summary becomes a restrictive or constraining issue for extra research according to author [16]. In addition to that, the researcher revealed models developed with extractive approach have better accuracy than that of abstractive approach from review of related works.

4.7 Study Design

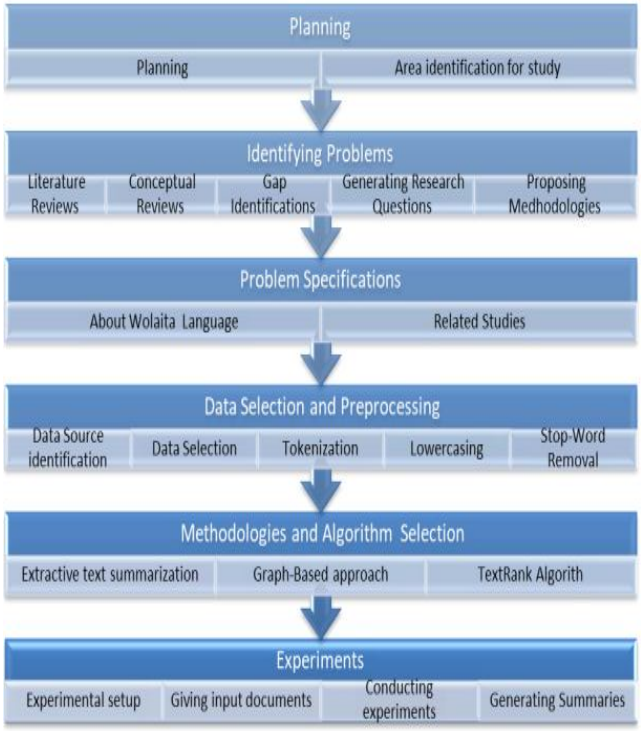

Figure 1.: Study Design

Figure 1 shows the design of study the researcher used to conduct text summarization for Wolaita language. It shows all the steps conducted by the researcher from problem identification to the final model which summarization given text by using TextRank algorithm which is one of graph based approach in the extractive text summarization method.

Shortly it design can be represented as;-

Planning -> Problem identification -> Problem specification ->Data selection-> Preprocessing -> Selecting design methodologies -> Selecting Algorithm -> Experiments -> Finally Generated Summaries.

Methodology part can be easily represented as

Text Summarization -> Extractive Approach - > Graph based approach -> TextRank Algorithm -> Summarization Model.

Summarized text as an input -> Summarization model -> Summarized text

5. ARCHITECTURE AND SYSTEM DESIGN

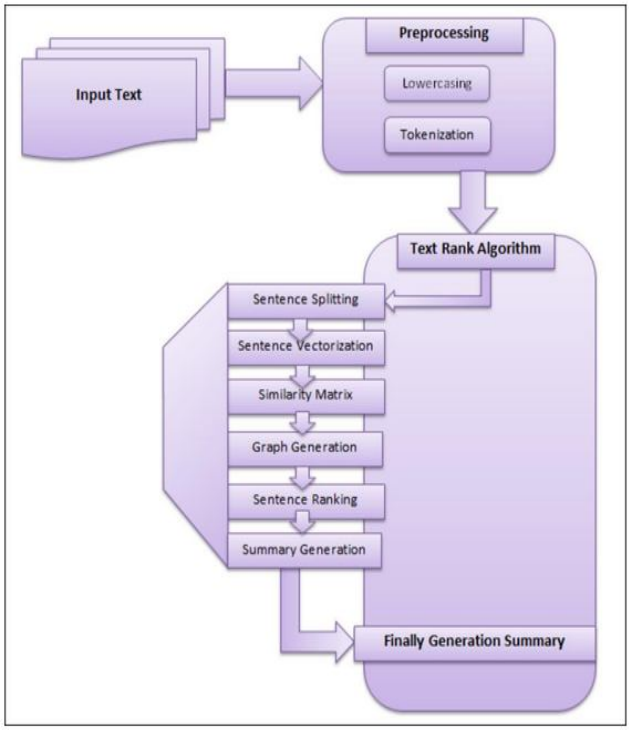

Figure 2: Architecture and System Design

Figure 2 shows the architecture and design of an algorithm and steps of entire summarization process.

6. EXPERIMENTS

In this study experiments were performed for totally 92 documents. Experimental findings for the documents containing maximum number of sentences and words; average number of sentences and words and minimum number of sentences and words were described for documents in this paper. In the discussion part, the researcher explained the findings for total documents by precision recall and f-measure by evaluating the system summaries with expert summaries. The researcher collected reference summaries from three experts for whole documents.

Table 3: Statistics of Data Corpus used for study

Roll No	Corpus Attributes	Values in Number
1	Number of documents	92
2	Maximum Number of Sentences per document	20
3	Minimum Number of Sentences per document	7
4	Average Number of Sentences per document	12
5	Maximum Number of words per document	825
6	Minimum Number of words per document	113
7	Average Number of words per document	360

Table 3 shows statistics for data corpus used for the study. It shows total number of documents used as an input text for the study, the maximum number sentences per document from input documents, average number of sentences from input document and minimum number of sentences for input documents. It also shows maximum, average and minimum number of words for documents used for study.

Table 4: Result of Experiment

Document	Input File			Summary			
	Length	#Sentences	#Words	Length	#Sentences	#Word	Reduction in %
Data13	1969	19	235	700	5	85	64%
Data33	1785	20	185	540	5	59	69.1%
Data80	1004	12	113	273	4	35	69%
Data70	360	7	44	160	3	20	57%

Table 3 shows the experimental result of the study. It shows key for input text, length of input file, number of sentences for and number of words for input test before summarization and after summarization. It also shows reduction in percent for summarized text.

Figure 3: Diagrams & Sentences Ranks of documents with maximum words

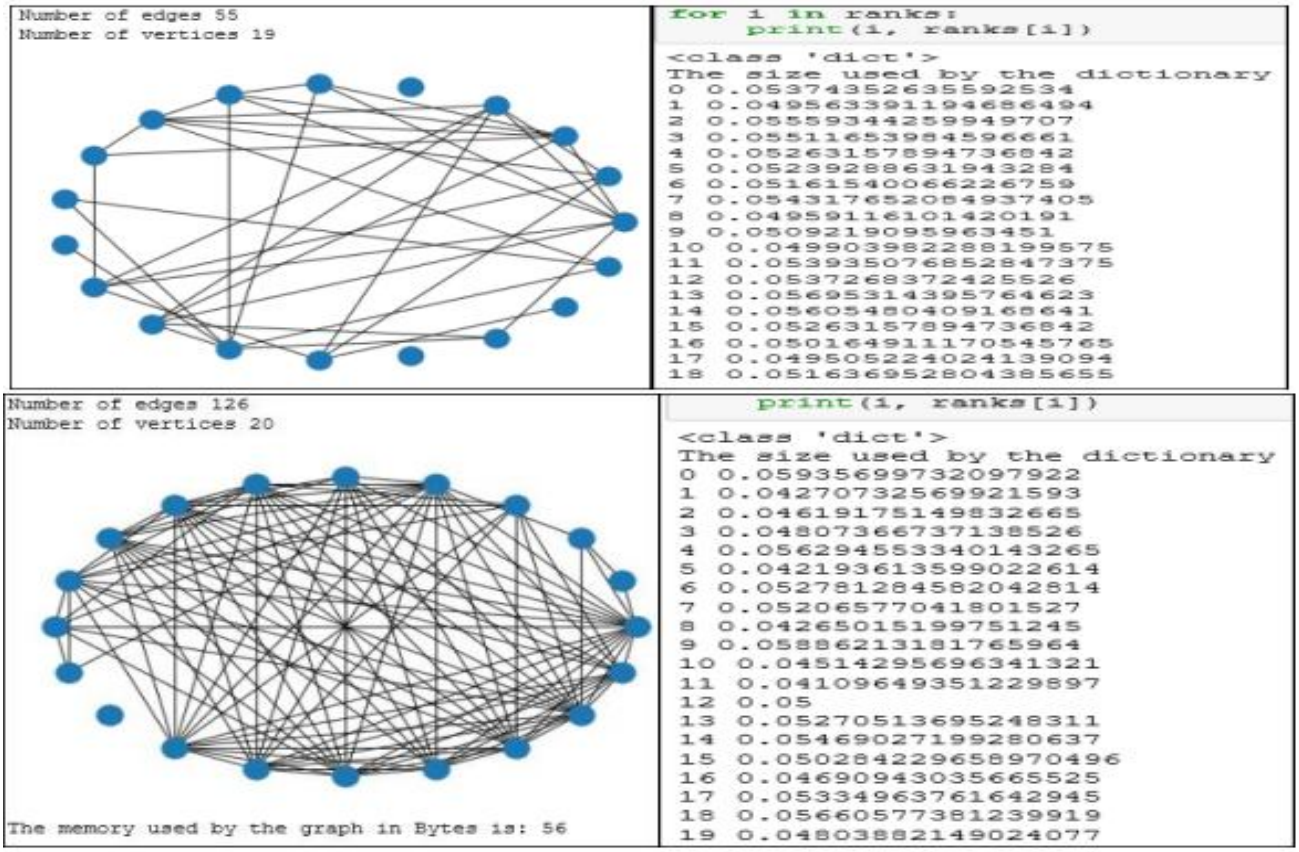

Figure 4: shows the diagrams & Sentences Ranks of documents with maximum words.

In the above diagram, the nodes or edges connected to more other nodes represents that such sentence is more related to the connected sentences meaning that there is more sentence similarity between such sentences. It also signifies that sentences have the more chance or probability to be included in the finally generated summary of the model. On the other hand the node which has no connection or less connection to the other node represents that the words in such kind of documents are almost not similar or there is less sentence similarity between such sentences. Plus such sentences has less chance to be included in finally generated summary by the model.

7. RESULT DISCUSSIONS

The main aim of this study was to develop text extractive text summarizer for Wolaita. As we can see from the findings of the above experiments in table 4, the researcher explained reduction of the input documents in percent with summarized documents. The findings were generalized as follows.

Table 5: Description of model generated summary

Experiments	Number of Sentences in input text	No of sentences in summary	Reduction in Percentage
1	19	5	64%
2	20	5	69.1%
3	12	4	69%
4	7	3	57%

It shows percentage of reduction for the documents. The maximum numbers of sentences are 20 and minimum number of sentences is 7 in the documents. Average number of sentences in the document is 12. All such sentences were ranked according to their scores in the dictionary. The system selected the sentences with maximum number of sentence scores and included them in the final summary. In the experiments, from 19 and 20 sentences, 5 sentences for each according to their sentence scores were selected and from the input document and included in the final summary. While forming the final summary, from selected sentences, the sentence with maximum sentence score becomes at the start and sentence with the lowest score becomes lastly in the final summary. Four and three sentences by the sentence scores were selected from the input documents with average and minimum number of sentences respectively and arranged by similar approach in the final summary. While we check the reduction of original document in percentage, it was 64%, 69%, 69.1% and 57% for experiments respectively and from these result, we can say input documents were reduced by 65% in the system generated summary. Accordingly, it has been done for all experiments. Generally the developed extractive text summarization model developed for Wolaita language better summarizes the documents used for summaries From these points we see that the developed extractive text summarization model performs summary with better performance. These could be justified in the performance evaluation in the next part with expert summarizations. In this experiment, while we check all features for the input text and the summarized text, the summarization model reduced the original text by 57%-69.1%. This was checked by giving 92 documents used for training the model and develop extractive text summarization model for Wolaita.

8. EVALUATION OF THE RESULT

To evaluate the performance of the system summary, we have collected manual summaries for documents used for training from three experts. Similar documents were given for the experts. Finally the system summary was compared with the expert summaries. ROUGE (Recall-Oriented Understudy for Gisting Evaluation), is basically of a set of metrics for evaluating automatic text summarization plus machine translation. This metric works by comparing system generated summary or translation against a set of expert or reference summaries. Reference summary is called human produced summary. The ROUGE Results were explained in terms of precision, recall and f-measures as given below.

$$\text{Recall} = \frac{\#\text{ of overlapping words of Manual and system summaries}}{\text{Total number of words in the manual summary}}$$

Most of the time, a machine generated summary could be extremely long, and it might capture all words in the expert summary. But, many words in the system generated summary might be useless and can make the summary unreasonably talkative. To handle this problem precision comes into effect. Precision essentially measures how much of the system summary was in fact relevant or needed. It is measures as follows. Precision = #of overlapping words of manual and system summaries Total # of words in system summary The precision feature is certainly crucial to generate concise summaries in nature. Therefore, it is continually best to calculate Precision and Recall then report the FMeasure. While summaries are in some way enforced to be concise via particular constraints, then Recall is considered to be used since precision is of less concern in this scenario. It is calculated as follows.

$$\text{F-Score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

In order to evaluate performance of the developed summarization model, the researcher used the precision, recall, F-measure metric as stated above. This evaluation metric was used for performance measurement because our summarization model produces extractive summarization. The summarization algorithm was evaluated for precision, recall and f-measure according to the above equations. To represent the performance of model the researcher computed the following steps:

1. Calculate the Precision, recall and F-measure obtained by summaries of the algorithm for all documents with respect to each reference or expert summary.

2. Sum up all of the resulting precision, recall and f-measure values and divide it by 92 × 3 = 276.

3. This gives the average precision, recall and f-measure attained by the algorithm with respect to the 92 documents and the three human expert evaluators.

Table 6: Average result of 92 input documents

No	92 documents used		
	Recall	Precision	F-M
	61.16%	60.69%	60.49%

Table 4 shows the average result of 92 documents by using ROUGE for three experts.

9. DESIGN ON TextRank ALGORITHM

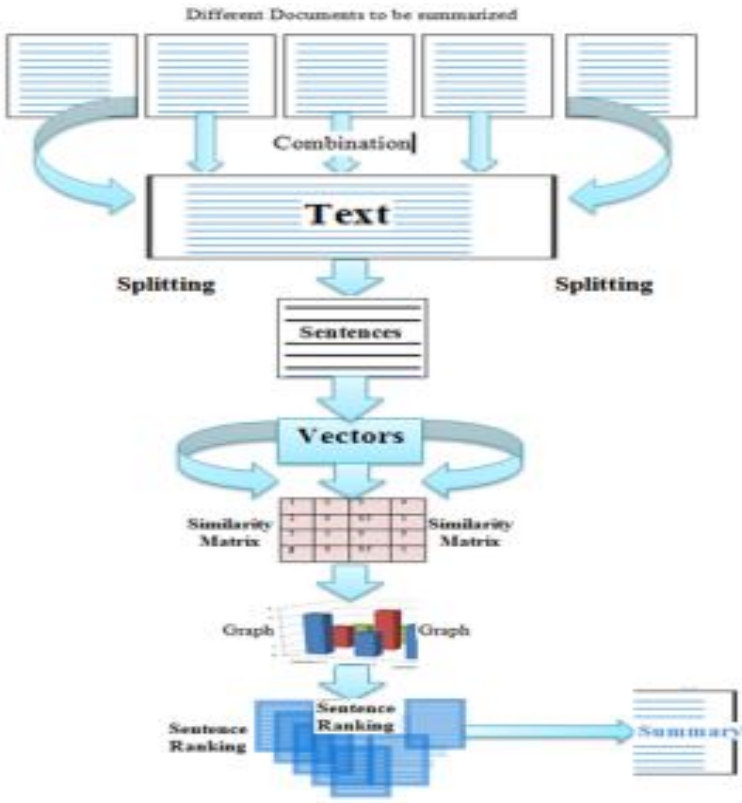

Figure 4: TextRank algorithm design

Figure 5 shows the design of TextRank Algorithm. It shows that TextRank can take text/document as an input, combine it as a single document and it into sentences vectors, generates a similarity matrix for each sentences, generate graph for sentences rank sentences according to their vector matrix and finally generate a summary for input document.

10. SIGNIFICANCE OF THE STUDY

Reading huge document wastes time and effort, rather than reading the entire article, dissecting it and separating the important ideas from the raw text which takes time and effort, it helps to get summary by short time with accurate information. It allows the user to read less data but still receive the most important information and make solid conclusions. Since summarizers work on linguistic models, they are able to summarize texts in most languages; therefore it can be used for other languages too. So stakeholders of this model can get short and precise information from languages they are not familiarized with and know their meanings by integrating it with other machine translation systems. It has also economic impacts, as it improves productivity as it speeds up the surfing process. Therefore various business organizations such as advertisers, massmedia owners, broadcast and radio stations can use this model in order to save their productive time. While doing so, they do not miss important ideas; it always mentions important ideas in sentence. Other organizations that perform activities related to text summarizations can use it for their intended purposes.

11. CONCLUSIONS

This study was designed to develop extractive text summarization for Wolaita language. In order to do this, the researcher reviewed various literatures and related to works. The researcher identified some gaps from review of related works and planned to put her contribution. In order to do that, the researcher performed the various objectives to accomplish the main goal or objective of study. The researcher has prepared applicable documents for Wolaita text summarizer totally 92 documents from primary level Wolaita language text books, explored extractive text summarization approach to address predefined problems, designed and developed extractive text summarization model for Wolaita language by using TextRank algorithm from graphbased extractive text summarization approach. The researcher evaluated the performance of text summarizer model by using ROUGE evaluation metrics and it has obtained promising result while compared with the previous works and finally reported result and recommended further research directions for interested future research scholars. The researcher addressed the problem statements with extractive text summarization approach by using graph based approach. From graph based approach, the researcher used TextRank algorithm to develop summarizer and finally evaluated summarization model by ROUGE evaluation metrics TextRank algorithm better summarized texts of Wolaita language than other summarizers and selected by the researcher. The researcher justified it with ROUGE evaluation metrics by comparing the system summaries with the expert summaries. Finally the accuracy and performance of developed model was expressed by recall, precision and f-measure i.e. 61.16%, 60.69% and 60.49% respectively. It shows that the developed extractive

text summarization model for Wolaita can better summarize texts of Wolaita language.

12. RECOMMENDATIONS

The researcher recommends the future interested researchers conduct this study with deep learning, artificial neural network (ANN) or convolutional neural network (CNN) to get the better results or model with the better accuracy compared to the findings of this study. Interested researchers can use the same documents and develop better summarizer model for Wolaita with PageRank and other graph based approaches like gensium approach. Most of articles and related works referred in this study and similar approaches conducted extractive text summarizations for the single documents. Therefore the researcher recommends the future researchers to conduct studies for multiple documents as multi documents summarizations for interested languages. In addition to that the researcher also recommends the future researchers to conduct text summarization for Wolaita language by using abstractive text summarization approach for both single and multi-document text summarizations. An interested researcher can also follow the same steps used in this study for multidocument text summarizations for Wolaita language.

13. ACKNOWLEDGEMENTS

We are thankful to the editorial boards and examiners of our article next to God almighty.

REFERENCES

[1] Ann Capestake, **Natural Language Processing**, Ann Capestake, Ed. USA, America: Ann Capestake, 2004.

[2] Kamal Sarkar and Sohini Roy Chowdhury Santanu Dam, "**A Study on Effect of Positional Information in Single Document Text Summarization**," International Journal of Cmputing Applications, vol. 16, no. No 1, pp. 29-37, January-June 2018.

[3] Josef Steinberger and Karel Jeˇzek,"**Text Summarization: An Old Challenge and New Approaches**," Springer, vol. 6, no. SCI 206, pp. 127- 149, 2010.

[4] Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutier-rez, and Krys Kochut. Mehdi Allahyari, "**Text Summarion Techniques**: Abrief Surver," CoRR, abs/1707.02268, 2017.

[5] Wen Xiao and Giuseppe Carenini, "**Extractive Summarization of Long Documents by Combining Global and Local Context**," University of British Columbia, Department of Computer Science, Vancouver, BC, Canada, V6T 1Z4, 2019.

[6] Guesh Amiha Birhanu and Wondwossen Mulugeta, "**Automatic Text Summarizer for Tigrinya Language** ," Addis Ababa University, Addis Ababa, MSc Thesis 2017.

[7] Shofi Ullah A. B. M. Alim Al Islam, "**A Framework for Extractive Text Summarization using Semantic Graph Based Approach**," ACM, no. ISBN 978-1-4503-7699-0/19/12., p. 1, December 2019.

[8] Elena Lloret, Rafael Mu˜noz, and Manuel Palomar Tatiana Vodolazova, "**Extractive Text Summarization**: CanWe Use the Same Techniques for Any Text?," Department of Language and InformationSystems,.

[9] K Selvani Deepthi, Shaik Ameen, Ravivarma G, M Mounisha Sai Teja Polisetty, "**Extractive Text Summarization for Sports Articles using Statistical Method**," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 1, 2 & 6, March 2020.

[10] Seyedamin Pouriyeh and Mehdi Assefi Mehdi Allahyari, "**Text Summarization Techniques**: A Brief Survey," vol. 3, no. 02268, p. 1, July 2017.

[11] Günes Erkan and Dragomir R Radev.,"**Graph-based lexical centrality as salience in text summarization**.," Journal of Artificial Intteligence, pp. 457–479., 2004. [12] Jiacheng Xu and Greg Durrett, "Neural Extractive **Text Summarization with Syntactic Compression**," 2019.

[13] Matej Gallo, "**Text Summarization by Machine Learning**," Masaryk University Faculty of Informatics, Masaryk, Msc Thesis Msc Thesis, 2016.

[14] Hans P. LUHN., "TheAutomatic Creation of LiteratureAbstracts," pp. 159-166, April 2016.

[15] Pauliina Anttila, "**Automatic Text Summarization**," University of Turku, Turku, Msc Thesis 2018.

[16] Mattias Gessesse Argaw, "**Efficient Language Independent Text Summarization Using Graph Based Approach**," Addis Ababa University, Addis Ababa, Msc Thesis 2015.

[17] Eyob Delele Yirdaw, "**Topic-based Amharic Text Summarization**," Addis Ababa University, Addis Ababa, MSc Thesis 2011.

[18] Melese Tamiru, "**AUTOMATIC AMHARIC TEXT SUMMARIZATION USING LATENT SEMANTIC ANALYSIS**," Addis Ababa University, Addis Ababa, MSc Thesis 2009.

[19] Mulugeta Getachew Regassa, "Topicbased Tigrigna Text Summarization Using WordNet," Addis Ababa University, Addis Ababa, MSc Thesis 2017.

[20] Fiseha Berhanu Tesema, "**Afan Oromo Automatic News Text Summarizer Based on Sentence Selection Function**," Addis Ababa University, Addis Ababa, MSc Thesis 2013.