2024-08-16

DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR þÿ A W N G I (Í š) N E W S D O C U M E N T

MULUKEN, TILAHUN KASSA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POST GRADUATE STUDIES

FACULTY OF COMPUTING

Program: -MSc in Computer Science

DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR AWNGI (አውኒ) NEWS DOCUMENTS

By: - MULUKEN TILAHUN KASSA

Advisor Name: Alemu Kumelachew (Ass.Prof)

BAHIR DAR, ETHIOPIA

August 16, 2024

DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR AWNGI (አውኒ) NEWS DOCUMENTS

By: - MULUKEN TILAHUN KASSA

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology in partial fulfillment of the requirements for the degree of MASTERS in computer science in the computing faculty.

BAHIR DAR, ETHIOPIA

2024

Bahir Dar Institute of Technology
School of Graduate Studies
Faculty of Computing
Request letter for proposal, progress, and final thesis presentation

Note.

- The presentation date and time requested by a student is according to the calendar of SRGS. In this case, the student will be notified within one week of receiving the request by the Advisor
- A student will be allowed to take a defense or research study presentation if he/she has completed a number of prior graduation requirements which include (1) **checking the similarity index of the thesis document** and (2) **preparing and submitting research article manuscript for publication to advisors, journals or conferences.**
- This form must be submitted by the student **4 weeks** before the presentation date to his /her advisor.
- Thesis corrections should be implemented by the student based on the examination committee's comments and recommendations.

Semester: - II Academic year: - 2024
Name of student: - Muluken Tilahun Student ID: - BDU1300618
Degree: - ☑ MSc ☐ MEng Program: - Computer Science
Email: - MulukenTilahun1802@gmail.com Phone Number: - 0938141665
Name of Advisor: - Alemu Kumelachew Academic rank: - (Ass. Prof)
Thesis title:- Design and Implementation of An Extractive Text Summarization System for Awngi (አውኒ) News Documents

I intend to take (select one from the given options) **Examination on** **Time**

☐ Thesis proposal defense	……../……../……..	…….
☐ Thesis progress defense	……../……../……..	…….
☑Thesis defense	……../……../……..	…….

Supporting documents for thesis defense (select one from the given options)

☐ Thesis proposal document

☐ Thesis progress report

☑ Thesis document

Checked by _____ date_____

_____ _____

Student's signature Advisor's Signature

Declaration

I hereby declare that the work presented in the thesis entitled, "**DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR AWNGI (አውኚ) NEWS DOCUMENTS**", is my original research work and that any sources of information, data, and ideas that have been used or referred to, have been properly acknowledged and cited.

As far as this thesis is concerned, I affirm that it has not been submitted for any other academic degree or qualification at another university. Considering the research conducted and results presented in this thesis are concerned, they are solely my work.

Moreover, I acknowledge that any assistance received during this research, including guidance from my supervisor(s), contributions from fellow researchers, or support from organizations or individuals, has been acknowledged in the Acknowledgments section.

Name: Muluken Tilahun

Date: August 16, 2024

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULITY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of student: - Muluken Tilahun Signature: - Date: - 9/2/2024

As members of the board of examiners, we examined this thesis entitled "DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR AWNGI (አውንጊ) NEWS DOCUMENTS" by *Muluken Tilahun*. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of science in "Computer Science".

Board of Examiners

Name of Advisor Alemu Kumelachew (Asst. Prof.)	Signature	Date 9/2/2024
Name of External examiner Dr. Martha Yifiru	Signature	Date Sep. 06, 2024
Name of Internal Examiner Dr. Debas Senshaw	Signature	Date 09/16/2024
Name of Chairperson Dagnachew Melese (Asst. Prof.)	Signature	Date 09/16/2024
Name of Chair Holder Kidist Meshesha	Signature	Date 06-01-17 E.C
Name of Faculty Dean Dr. Tesfa Tegegne		06-01-17

ACKNOWLEDGEMENTS

First and foremost, praises and thanks to God, the Almighty, and St. Mary the mother of Jesus for their showers of blessings throughout my research work.

I would like to express my deep and sincere gratitude to my research advisor Mr. Alemu Kumilachew (Asst. Prof) for allowing me to do research with him and providing invaluable guidance throughout this research. He showed me the methods for conducting research and presenting the results as plainly as possible. Working and studying under his direction was a wonderful honor and privilege. I am appreciative of everything he has done for me. I would also like to express my gratitude for his friendship, sensitivity, and good humor.

I am extending my thanks to the Bahir Dar University's computing faculty dean, and computer science staff members for their invaluable help in each journey of my research work. My sincere gratitude goes to Injibara University who let me work on my research paper by providing free internet and library access. Thanks so much, for all their invaluable help.

Finally, I am extremely grateful to my parents for their love, prayers, care, and sacrifices in educating and preparing me for my bright future.

ABSTRACT

The World Wide Web is a tremendous source of knowledge. With so much information available on the internet, humans are experiencing a problem of information overload. Therefore, a significant and automatic tool is required to convert lengthy documents into concise forms by extracting relevant information. Summarization involves condensing a body of data to create a concise summary that captures the most pertinent details from the original text. Extractive and abstractive summarization approaches are the two categories.

The aim of this paper was the design and implementation of an extractive summarization system for Awngi news documents. Amhara Media Corporation website having a newspaper for cherbewa, and a hard copy collected from the library and shop was the source of data. We collected a carefully selected dataset of 213 Awngi documents, preprocessed the data, and built a refined summarization model. The experiment was tested with three extractive summarization techniques (LSA, Text rank, and TF-IDF). The summary generated by LSA overperformed the other. The suggested model performs well in ROUGE evaluations, scoring an F1-score of 43.1% for ROUGE-1; an F1-score of 45.3% for ROUGE-2; and an F1-score 59.0% for ROUGE-L at the extraction rate of 40%. Furthermore, the proposed approach's (LSA) summary provides a cosine similarity of 84.55% with a human reference summary at the extraction rate of 40%. Our study gives an important insight into boosting summarizing performance in this particular linguistic setting and shows how successful extractive summarization is for Awngi text documents.

Keywords: - Summarization, Extractive, Awngi, LSA, TR, TF-IDF

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

(AI): - Artificial Intelligence

(ATS): - Automatic Text Summarization

(BOWs): - Bag of Words

(CSV): -Comma Separated Value

(HRS): - Human reference Summary

(LSA): - Latent Semantic Analysis

(MDS): - Multi-Document Summarization

(NLP): - Natural Language Processing

(NLTK): - Natural Language Tool Kit

(QA): - Question Answering

(ROUGE): -Recall-Oriented Understudy for Gisting Evaluation

(SDS): - Single Document Summarization

(SSD): - Solid State Drive

(TF-IDF): -Term Frequency Inverse Document Frequency

(TR): - Text Rank

(TV): - Television

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1. INTRODUCTION

1.1 Background

As a research field, natural language processing (NLP) received significant attention in the 1980s and 1990s. As a result of this period, researchers developed techniques and algorithms for enabling computers to interpret, understand, and generate human language with great speed. During these decades, the surge in interest in natural language processing laid the groundwork for the language processing capabilities we are seeing today in artificial intelligence and digital assistants.

In contrast with the invention of NLP systems, automatic text summarization was invented in the late 1950s, when there was a special interest in automating summaries for the creation of technical documentation abstracts (Saggion & Poibeau, 2013).

Text summarization is an active area of research in the field of natural language processing (NLP). It is a technique for creating summaries of text documents by taking the key details out of the entire document (T. Sri et al., 2017). A successful text summarization system should comprehend the entirety of the text, rearrange the information, and provide cohesive, educational, and astonishing summaries to communicate the key points of the original text (Kerui et al., 2020).

Based on the way of transforming a text document summary, there are two main summarization techniques: extractive summarization, where a representative set of sentences is selected from the whole document, and abstractive summarization, where the content of the summary is different from the original document's content. More of the research in text summarization deals with an extractive type of summarization since it does not need any kind of linguistic knowledge. Text summarization methods can also be single-

document or multi-document based on the number of inputs. Dealing with various inputs is more difficult since crucial information is spread over a collection of potentially disparate papers. Cohesion and reference resolution might be issued in multi-document summarization since sentences are combined from many documents (Moradi & Ghadiri, 2019). Generic and user-oriented summaries are another category. In addition to the above types of classifications, there are others, such as supervised and unsupervised (machine learning), informative, indicative, and so on.

One of the main issues was evaluating text summaries that were automatically created. Even though there are various evaluation metrics, the following are the major ones: *Evaluation by sentence co-selection*, precision, and recall are the measures for the co-selected sentences. This measure needs the right extract or a reference to make a comparison; more of the reference summaries are from human experts. *The context-based method* compares two documents more closely than only by comparing their sentences. The fundamental technique involves using the cosine similarity measure to determine how similar the entire text document and its summary are (Steinberger & Ježek, 2004). Relevance evaluation, task-based evaluation, and evaluation based on latent semantic analysis similarity of the main topic are such different evaluation techniques.

It is still necessary to have an automated system in place to help extract valuable information from a vast repository of text documents that are accessible through social media and the Internet (Kerui et al., 2020). It would be extremely helpful to have an automated system that sifts through this wealth of information intelligently and identifies relevant, meaningful insights. As a result, knowledge extraction could be more efficient, decision-making could be improved, and textual data that is now widely available could be utilized to its fullest potential.

In this study, the researcher proposed an extractive text summarization system for Awngi news documents. Using an interactive user interface, the system extracts concise information from a longer source document. individuals can absorb key information from extensive textual materials quickly and efficiently without having to read all of the original material. When working with large, complex documents, summarization is an important feature that enhances productivity and information comprehension. With this integrated approach, users can effectively extract the most important points from lengthy documents, which saves them time and cognitive effort.

1.2 Statement of the problem

Textual data in the form of digital documents quickly adds up to massive volumes of information. According to (Garbade, 2018) the total amount of digital data around the world is supposed to hit 4.4 zettabytes in 2013 to 180 zettabytes in 2025. During the era of information explosion, the ability to process unstructured documents efficiently has become increasingly crucial for extracting valuable insights and driving informed decisions. Unstructured documents, however, are often problematic, as they lack organization, ambiguity, and difficult information extraction, which hinders the seamless extraction of key information. As a result, internet-age users waste a significant amount of time exploring and reading to find and extract a few topics or items of interest in morphologically rich languages such as Awngi. An extractive summarization has been extensively studied in the literature. A research study by (Guadie et al., 2021) introduced an Amharic news text summarization for news items posted on social media specifically on Twitter and Facebook. As part of their proposed approach, they calculate the similarity between the two pairs of posted documents. Using the K-means algorithm, then cluster the documents based on their similarity results. Finally, summarize each clustered post using TF-IDF algorithms, which are statistical methods for ranking the documents based on their frequency. Much research

has been done on foreign languages such as Hindi (., 2016; M. Gupta & Garg, 2016a; Krishnakumar et al., 2022; Taunk & Varma, 2022).

(Dinegde & Tachbelie, 2014; Tashoma et al., 2020a) are the researches that were done for Afaan Oromo text summarization. A single document text summarization done by (Li et al., 2016) is also an extractive text summarization that was done for Tibetan a language spoken in the Himalayas. Text rank and LexRank were the techniques that were employed.

Although much research has been done in Extractive text summarization for foreign languages like English, and Hindi and for local languages such as Amharic, Afaan Oromo, and Tigregna, to the best of the researcher's knowledge, no research has been done on the Awngi language. The distinctive qualities and nuances that are inherent in Awngi's writing may not be well captured by current generic text summarizing techniques, leading to poor summary results. Moreover, there is a literature gap that adds a body of knowledge for fellow researchers in this language domain. Providing a solution to this issue is extremely important as it increases the usability and effectiveness of extracting important information from the Awngi text document, allowing for more informed decision-making and an overall better user experience.

As a result, the aim of the study is to design and implement an extractive summarization system for Awngi news documents. In the end, this study will answer the following research questions:

- Which extractive text summarization technique outperforms the other?
- What is the optimal extraction rate for producing the most concise yet informative summaries of Awngi news documents?
- How to design and implement a text summarization system for Awngi news documents?

- How to evaluate the performance of generated summaries with human reference summaries?

1.3 Objective of the study

1.3.1 General objective

The General objective of this research work was to develop an effective and efficient extractive text summarization system capable of generating concise yet comprehensive summaries of Awngi news documents.

1.3.2 Specific objectives

- To study relevant literature on the area of text summarization and Awigni language.
- To prepare the Awngi text document dataset.
- To develop a text summarizer model for the Awngi news document.
- To design and implement a prototype for an Awngi news document summarization system.
- To evaluate the performance of the generated summary.

1.4 Scope of the study

The scope of this study was limited to the design and implementation of an extractive text summarization system for Awngi (አውኒ) news documents. The concerned tasks include literature review, problem identification, data collection, modeling, prototyping, and evaluation. The public website: https://www.ameco.et/category/ቸርቤዋ magazine and hard-printed magazines from shops and libraries were the sources of data. The study was limited to Awngi text documents only. An extractive summarization approach was employed which preserves the wording and structure of the original document.

1.5 Significance of the study

It tackles the peculiar difficulties and traits of Awngi's text, resulting in more accurate and pertinent summaries for this specific area. This study helps conserve time and resources by automating the summarizing process for the Awngi language rather than manually reading and examining extensive papers. The capacity to provide succinct summaries lowers the work needed to extract important information, improving the efficiency of information consumption. The results of this study have applicability in a variety of text-based Awngi applications. The immediate beneficiaries of the study are journalists and report generators.

Future studies in the area of Awngi text summaries can build on the research done in this paper. It may motivate experts to investigate new methods, enhancements, and uses specifically for Awngi text, resulting in more development in the area.

Furthermore, It can be used as a reference for fellow researchers in academics with an interest in Awngi text summaries or similar fields.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Introduction

A literature review is a critical and thorough examination of the body of knowledge already available on a certain subject or research issue from published books, papers, research projects, and other sources. Its purpose is to provide an overview, assessment, and synthesis of the state of the art in a certain field. In this chapter, the Awngi language, its alphabets, phonology, morphology, syntax, and semantics are reviewed. Moreover, the definitions, their importance, the classifications, the approaches or techniques, the preprocessing steps, evaluation metrics, and applications of an automatic text summarization system are discussed in detail.

2.2 The Awngi language

Language serves as a medium of communication for human beings. it has properties such as being human, transferability from one generation to another, birth, development, and death are among the different properties of every language on its own.

Ethiopia is one of the African countries that have more than 86 languages and more than 200 dialects spoken. Oromos, Amhara, and Tigrayans are among the largest ethnic groups that are spoken in Ethiopia. Ge'ez is one of the ancient and original languages of Ethiopia that has served as a written and spoken language for the Ethiopian orthodox church until now. Amharic and Tigrigna were derived from this ancient language (Languages of Ethiopia, 2021).

Awngi is part of the Cushitic language where the alphabet used is derived from the Ge'ez script. As compared to Ge'ez scripts having 33 letters each of which denotes 7 characters,

which constitutes 231 characters as a total, the Awngi language has 26 letters each of which denotes 7 characters a total of 182 characters.

Awi nationality administration is one of the three autonomous nationality zones that are found in the Amhara regional state, it is also called the Agew-Awi Zone. The landscape in the Awi province consists of lowlands, semi-highlands, and highlands.

The population of Awi is estimated to be 1,159,385 (Shiferaw, 2015). The inhabitants are predominantly the Awi ethnic group but with diverse dialects. Awi is a sub-group of the Agew people whose historical contribution dates back to the Zagwe Dynasty. Commonly spoken languages in the zone are Amharic and Awngi.

Agaw (አገው) is a collective name for four Cushitic-speaking ethnic groups in Ethiopia and Eritrea. The Bilen are from Eritrea, the Kimant (ቅማንት) are from Gonder, the Xamtanga are from Wag-Sekota (ዋግ-ሰቆጣ), and the Awngi are from Gojjam (Agajie, 2020). Awngi is a Central Cushitic language and the majority of speakers can be found in the Agew Awi Zone of the Amhara Region (Ager, 2021). It is also spoken in Metekel (መተከል), Dangur(ዳንጉር), and Quara (ቋራ), Gonder.

Awngi language was started being taught at primary school in late 1989, in five schools with 10 teachers and 263 students. In 2007 the number of schools increased to 253 with 3011 teachers and 117386 students, this shows how the language is developing tremendously(Shiferaw, 2015). Nowadays, the language has media coverage in Amhara Media Corporation, a radio program, and a newspaper publication such as “ቺርቤዋ”.

2.2.1 Awngi Alphabets (አውኚኩ ፊደልካ)

Table 1: - Awngi alphabets

በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ

ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ		ወ	ው	ዊ	ዋ	ዌ	ው	ዎ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ		ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ		መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ		የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
አ	ኡ	ኢ	አ	ኤ	እ	ኦ		ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ		ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ		ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ		ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ		ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ		ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ

2.2.2 Pronouns (ስም ባና)

A pronoun is a word that is used instead of a noun. A pronoun is very much used in our day-to-day conversation.

Table 2: - Pronouns

	Subject pronouns	
English	**Awigna**	**Amharic**
I	አን	እኔ
We	እኖጂ	እኛ
You (M)	እንት	አንተ
You (F)	እንት	አንቺ
She	ኚ	እሷ
You (Pl)	እንቶጂ	እናንተ
He	ኚ	እሱ
They	ጋጂ	እነሱ

2.2.3 Possessive pronouns

Possessive pronouns are used to show the possession of the noun or ownership. They are also used to replace a noun or noun phrase to avoid repetition.

Table 3: - Possessive pronouns

English	Awigna	Amharic
My	ይው	የኔ
Our	እንው/እኖጂሱ	የኛ
Your (M)	ኩው	ያንተ
Your (F)	ኩው	ያንቺ
Her	ኺው	የሷ
Your (pl)	እንቶጂሱ	የናንተ
His	ኺው	የሱ
Their	ጆጂሱ	የነሱ

2.2.4 Preposition

A preposition is a word that is placed before a noun or pronoun. In the Awigna language prepositions come after a noun/pronoun.

Table 4: - Prepositions

English	Awigna	Amharic
In	አኽ	ውስጥ
Inside	አኽዳ	ከውስጥ
Outside	ሴግዳ	ከውጭ
Above	አምፐ	ላይ
Below	ኩክሪ	ታች
Under	ስርዳ	ከስር
Behind	እንግርዳ	ከኋላ

English	Awigna	Amharic
In front of	ፍንዳ	ፊትለፊት
Beside	ናቕት	አጠገብ
Opposite	ፊሪስዊ	ተቃራኒ
Next to	ሲፋማ	ቀጥሎ
Of	. . . . . ው	የ
For	. . . . . ስ	ለ
By	. . . . . ዳ/ስ	በ
From	. . . . .ዴስ	ከ
To/towards	. . . . ሾ	ወደ
Around	ዙሪዳ	በዙሪያ
With	. . . . ሊ	ጋር

2.2.5 Conjunctions

In sentence formation, conjunctions have great importance in connecting words, clauses, and phrases. Among the different types of conjunctions used the following are listed: -

Table 5: - Conjunctions

English	Awigna	Amharic
And	እስታ	እና
But	ያኼስጉ	ግን
Or	አኾኪ	ወይም
Because	ምክንያትኪ	ምክናየቱም

2.2.6 Determiners

Determiners are words that introduce a noun and the quantity of a noun that is used in a sentence.

Table 6: - Determiners

English	Awigna	Amharic
This	እን	ይህ
These	እኒ	እነዚህ
That	አን	ያ
Those	አኒ	እነዚያ
Here	እንዳ	እዚህ
There	ኺዳ/አንዳ	እዚያ

2.3 Phonology

It is the study of the patterns of sounds that can be found in a language and across languages. Verbally, phonology describes how speech sounds can be categorized in the brain and communicate meaning.

2.3.1 Consonants

There are twenty-nine consonant phonemes in Awngi, five of which are labialized.

Table 7: - Awngi Consonant Phonemes

	labial	**alveolar**	**Palato-velar**	**uvular**	
Voiceless plosives	p	t	k	q	plain
			kʷ	qʷ	labialized
Voiced plosives	b	d	g	ɢ [ʁ]	plain
			gʷ	ɢʷ [ʁʷ]	labialized
Voiceless affricates		ʦ	ʧ		
voiced affricates		ʣ [z]	ʤ		
fricatives	f	s	ʃ		
post-stopped fricatives		st			

nasals	m	n	ŋ		plain
			ŋ^{w}		labialized
lateral approximant		l			
vibrant		r [ɾ]			
approximant	w		j		

2.3.2 Vowels

Awngi has 6 vowel phonemes.

Table 8: - Awngi vowel phonemes

	Front	Central	Back
Close	i	(ɨ)	u
Non-close	e	a	o

2.4 Morphology

Morphology, within the field of linguistics, examines the structure of words, the rules governing their formation, and their interrelation within a particular language. The majority of morphological studies focus on the word structure through morphemes, the minimal meaningful language units. These morphemes encompass roots that stand alone as words, as well as categories like affixes, which are bound to other words. It also examines the behavior of words as parts of speech and their inflection to denote grammatical categories such as number, tense, and aspect. Moreover it explores concepts like productivity, which pertains to the ways in which speakers coin new words in particular contexts, a process that changes throughout the history of a language.

2.4.1 The nominal inflection

Nominal inflection is the process of altering the form of a noun, pronoun, adjective, or other nominal word to express grammatical categories like case, number, and gender.

2.4.1.1 Gender

In its singular form, the Awngi language distinguishes between two genders: masculine and feminine. In humans and domesticated animals, gender distinction signifies the differences between sexes.(Hetzron, 1978)

For example: - mulíqisí ‘monk’/ moleqésá ‘nun’, sén ‘brother’/séna ‘sister’. There are certain suppletive terms that refer to sex, e.g., aqí ‘man’/Ɣuna ‘woman’, kormí ‘stallion’/ bazrá ‘mare’ (fírisí ‘horse’). Most words denoting things have a fundamental gender of masculine. The use of the feminine has an affective, diminutive, derogatory, or sometimes caritative connotation. In many instances, the use of the feminine form can slightly alter the meaning compared to the masculine form. e.g., árfí ‘month’/ árfá ‘moon’, amét ‘year’/améta ‘next year’.

2.4.1.2 Number

Awngi language distinguishes only between singular and plural forms, without gender distinction in the plural. A plural noun can represent the pluralization of masculine, feminine, or both genders simultaneously. The most common plural marker is "-ka," which is attached to the final consonant of the stem, except in compounds. In the noun class where the masculine ends in -i, the bare stem form may also be used for the plural, e.g., dəƔʷari ‘donkey’/pl, dəƔʷarká or dəƔʷar.

2.5 Syntax

During syntactic analysis, words and morphemes are combined into larger units such as sentences and phrases. Sentences in English follow a subject-verb-object word order, which ensures syntactical accuracy.

Conversely, the Amharic and Awngi languages have subject-object-verb (SOV) grammatical patterns.

For example, "Kebede ate bread." (English)

"ከበደ ዳቦ በላ።" (Amharic)

"ከበደ ቱሼ ኹኻ።" (Awngi)

2.6 Semantics

In natural and artificial languages, semantics deals with meaning. It is the study of how language users create meaning via the use of linguistic structures, as well as how readers and listeners comprehend and interpret that meaning in various settings.

2.7 Why do we summarize texts?

Textual data in the form of digital documents quickly adds up to massive volumes of information. The vast majority of these papers are unstructured, meaning they are unconstrained text that has not been categorized into typical databases. Because of the lack of standards, processing papers is a rudimentary effort. As a result, implementing automatic text analysis tasks has become incredibly challenging. Automatic text summarizing (ATS) can assist in digesting this ever-increasing, difficult-to-handle amount of data by condensing the text while retaining significant information (Juan-Manuel et al., 2014).

There are various compelling arguments in favor of document summarizing by machine. Here are just a few of them:

1. Summaries help you read faster.
2. Summaries make the choosing process easier while researching documents.
3. Indexing is more effective with automatic summarization.
4. Human summarizers are more prejudiced than automatic summarizing techniques.
5. Because they give individualized information, personalized summaries are important in question-answering systems.
6. Commercial abstract services can enhance the volume of texts they can handle by using automatic or semi-automatic summarizing techniques.

2.8 Definitions of text summarization

The most apparent technique to shorten a paper is to use summaries. The job of creating a succinct and fluent summary while keeping vital information content and overall meaning is known as automatic text summarization (Allahyari et al., 2017). Additionally, the authors also defined Automatic text summarizing as a difficult task, as humans summarize a piece of text, they normally read it completely to have a thorough comprehension of it before writing a summary highlighting its important points. since computers lack human language and understanding, automated text summarization is a complex and time-consuming operation.

Text summarization is a well-established field of study whose primary purpose is to investigate and create summarization methods capable of extracting high-quality information from enormous document collections (Fiori, 2019).

Based on the type of input document(Torres Moreno, 2014) explained text summarization as the transformation of a source text into a reduced, shorter version that retains all key information. The input for single document summarization is just one document, and the summary is derived from it (Wazery et al., 2022).

Text summarization is a natural language processing approach that creates a small, crisp version of a huge textual material (Zaki et al., 2020). “Automatic text summarization is a subfield of Natural Language Processing (NLP) that aims at producing precise and non-redundant text aided by machine learning techniques” (Pattnaik & Nayak, 2019)

2.9 Classifications of text summarization

Different sets of criteria can be used to categorize summaries, based on input type the summarization process can be either single-document or multi-document summarization.

2.9.1 Single document summarization

Single document summarization is a summarization process where only one or a single document is taken as input to produce a summarized text. The automatic development of an abstract for a scientific paper is an example of a single-document summary problem. The abstract should cover all of the key topics discussed in the report, regardless of where they appear in the text.

2.9.2 Multi-document summarization

Whereas in multi-document summarization the input is from various multiple documents and all of these papers' information should be included in the summary. The news summarizing job is a prominent multi-document summary issue in which the aim is to summarize a group of news stories published in multiple newspapers and covering the same themes.

The other angle where the summarization process can be classified is based on output type, it can be either Extractive or Abstractive summarization (Wazery et al., 2022).

2.9.3 Extractive Summarization

The original sentences from the input material are used to create an extractive summary. Sentence extraction, statistical analysis, and machine learning approaches can all be used to create such summaries (Azhari & Jaya Kumar, 2017). (Joshi et al., 2021) also defined as Extractive summarization is a method of extracting delineative paragraphs or phrases from the original text and integrating them into a smaller document than the original.

The significance of the sentence is measured based on the statistical or linguistic features of the sentence. Content word (Keyword), title word, sentence placement, sentence length, proper noun, upper-case word Cue-phrase, Biased term, Font based, Pronouns, Cohesion between sentences, sentence-to-centroid cohesion, Discourse

analysis, and the occurrence of non-essential information are the most typical features utilized for extractive summarization.

Extractive summarization algorithms are further separated into supervised and unsupervised approaches. Supervised techniques use summarizing as a classification issue, categorizing document sentences into two groups: in-summary and not-in-summary. This study focuses on the extractive text summarization method as it is based on selecting an informative and indicative summary of the whole document, and it does not require any linguistic knowledge.

2.9.4 Abstractive Summarization

Abstractive summarizing is a type of summarization in which linguistic techniques are used to construct a reduced summary of a text. This type of summarizing is more efficient than extractive summarization since it may create new sentences that convey the most important information from the source (s) (Joshi et al., 2021).

On the other hand, (Pattnaik & Nayak, 2019) stated that the abstractive summarization technique creates summaries that are made up of words and phrases that describe the original document's information content. The output summary obtained from the abstractive summary is not a direct copy of the original document and they are harder to implement because they need deep linguistic knowledge.

According to the function, text summarization can be classified as either indicative or informative summary.

2.9.5 Indicative Summarization

The themes mentioned in the source document are summarized in an indicative summary. It's similar to a table of contents. Only the most significant notion in the text is presented in an indicative summary system. An indicative summary provides a high-level overview of the text's themes. This form of summary aids the user in deciding whether or not to

continue reading the material. The average length of this type of summary is 5 to 10% of the original content (S. H. B. Sri & Dutta, 2021).

2.9.6 Informative Summarization

An informative summary attempts to represent the source text's content, maybe elaborating the reasoning. It is a condensed version of the original document. Informative summaries are more difficult to write than indicative summaries because they require the source text's information to be properly understood, generalized, organized, and synthesized (Juan-Manuel et al., 2014). The informative summary approach includes every facet of the primary text. The informative summaries are around 20 to 30 percent of the original material in length (S. H. B. Sri & Dutta, 2021).

On the other hand, based on context or content, summarization can be categorized into three categories generic, query-driven, or domain-specific summaries.

2.9.7 Generic summarization

A document summary that bypasses the information demands of users. Generic summaries have no perspective on the subject and treat the document as a single text, so all information is treated equally.

2.9.8 Query-based summarization

It is a type of summary generated based on the information requirements of the user or user inquiries. The answers for this type of summarization are minimal with a limited number of words and phrases. This approach may be based on single or multi-document input types.

2.9.9 Domain-specific summarization

These types of summarization processes are based on a specific field or domain of study such as health, law, medicine, and the like (Bhattacharya et al., 2021; S. R. Patil, 2011; Reeve et al., 2007).

2.10 Approaches or techniques

Six approaches or techniques for text summarizing have been identified from the literature collected over the previous ten years: fuzzy-based, machine learning, statistics, graphics, topic modeling, and rule-based (T. Sri et al., 2017).

2.10.1 Fuzzy-based

It is the most popular approach as it avoids data inconsistencies. Instead of using the standard true or false (1 or 0) Boolean logic, fuzzy logic uses degrees of truth, which is more in line with human reasoning. The membership function and fuzzy rules used in the architecture of the fuzzy system have a significant impact on performance. Based on the features included in each sentence and the rules established in a knowledge base, a value ranging from zero to one is derived for each sentence in the output (Andhale & Bewoor, 2017). After all, determining ambiguity entails the function of humans. Fuzzy systems operate by using a variety of inputs from different attributes or indexes. The fuzzy inference system is then provided with the score of each feature as input for use with the IF-THEN rule of human knowledge. Most of the time this approach is used to produce extractive summaries.

2.10.2 Machine learning

It is the favorite technique in this modern era where the training set of data is given to the model while splitting some data for testing the performance of the developed model. The machine learning techniques improve the performance of the model from experience without being explicitly programmed.

As part of the machine learning approach, text summarization involves the following methods. SVM (Support vector Machine), K-Means, Naïve-Bayes, and deep learning. According to the paper by (Dingare et al., 2022) they describe, put into practice, and contrast some unsupervised machine learning methods, such as k-means clustering, latent

Dirichlet allocation, and latent semantic analysis. The paper by (Mattupalli* et al., 2020) employed A deep-learning approach known as Long Short Term Memory (LSTM) on the CNN daily mail dataset for extractive summarization.

2.10.3 Statistics

Statistics are frequently used to determine a feature's score or weight. For instance, a statistical method called TF-IDF may be used to identify the frequency of words, determine keywords, and determine the similarity that emerges. The final score of the sentence selected in the summary is then extracted or determined using a machine learning or fuzzy-based technique using the findings of the statistical approach as input.

A journal paper by (V. Patil et al., 2020) proposed a statistical approach for multiple document summarization with the major goal of providing users with additional contextual and summary information to help them find results more quickly by condensing collections of relevant Web pages.

2.10.4 Graph-Based

In essence, graph-based ranking algorithms determine the significance of a vertex inside a graph using data derived from the network structure. If two sentences have a semantic relationship, they are connected by an edge, and the weight of the edge is determined by the relationship. The ranking of graph vertex relevance is determined by a graph-based algorithm. High cardinality vertex sentences are regarded as significant sentences and are included in the summary. Neither the graph-based technique nor the main knowledge for summarizing calls for in-depth language expertise.

The Google Internet search engine uses a link analysis method called PageRank that gives each element of a set of hyperlinked documents, like the World Wide Web, a numerical weighting. PageRank, which is Google’s trademark, was named after one of the

co-founders of Google Inc. and a well-known American computer scientist and businessman Larry Page.

2.10.5 Topic Modeling

Topic modeling text summarization involves a document having various topics.

The subjects in the original material are found using topic modeling methods like LDA. The creation of text clusters is then done using these subjects. Significant sentences from the source material are included in the clusters. Each cluster would be related to the pertinent themes that had been chosen. Usually, to expand coverage and aid in the summarizing of the source content, this summarization technique assigns phrases in the document to multiple selected categories (Issam et al., 2020).

2.10.5.1 Latent semantic analysis

As a theory and method of meaning extraction, LSA brings together researchers from computer science, information retrieval, psychology, linguistics, cognitive science, information systems, education, and many other related fields to analyze word usage patterns to determine meaning (Joy Winnie Wise et al., 2024). This technique in Natural Language Processing (NLP) uncovers the latent structure of a text collection using Latent Semantic Analysis (LSA). In addition to reducing dimensionality, it is also used to discover relationships among terms and documents.

Based on the principle that words that occur in the same context have similar meanings, Latent Semantic Analysis utilizes the mathematical technique of Singular Value Decomposition (SVD) to identify patterns of relationships between the terms and concepts. SVD is one of the dimensionality reduction techniques. i.e., if matrix A is factored into three matrices, then matrix A will look like this: -

$$\boldsymbol{A = U \sum V^T}$$

A is a matrix with dimensions **m x n**. **U** is an orthogonal matrix with dimensions **m x m**. ∑ is a m x n diagonal matrix, while **V** is a **n x n** orthogonal matrix. **U** is known as a left singular vector, ∑ is a singular value or eigenvalue, and **V** is the right singular vector.

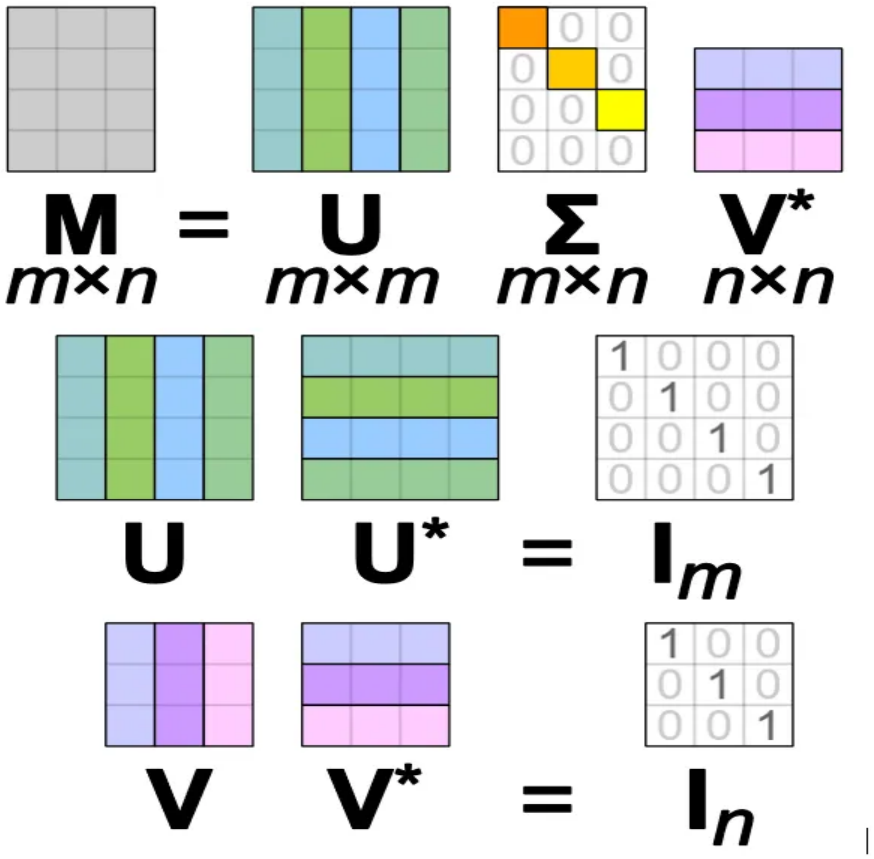


Figure 1: - Singular value decomposition visualization (Document Summarization Using Latent Semantic Indexing | by Srinivas Chakravarthy | Towards Data Science, n.d.).

2.10.6 Rule-Based

One of the first NLP techniques is the rule-based approach, which analyzes and processes textual input according to established language rules. Applying a certain set of rules or patterns to capture particular structures, extract information, or carry out activities like text categorization and other similar ones is known as a rule-based method. Pattern matching and regular expressions are two typical rule-based methods.

The scholars of Bangladesh (Protim Ghosh et al., 2018) proposed a rule-based extractive text summarization technique for Bangla news documents. For the first time, a

graph-based phrase grading feature was included in this suggested method for summarizing Bangla news documents.

The authors (M. Gupta & Garg, 2016b) also developed a text summarization system using a rule-based approach to the Hindi language while eliminating dead phrases and deadwood. It is based on the extraction of relevant information regardless of the semantics or meaning of the entire document.

2.11 Pre-processing steps

Text summarization is one of many natural language processing (NLP) operations that need preprocessing. It entails preparing unprocessed text data for analysis and further processing by cleaning and converting it. Here are a few typical preprocessing methods:

2.11.1 Text cleaning

This pre-processing step involves cleaning up the text document from any unnecessary or distracting parts, such as special characters, HTML tags, non-alphanumeric letters, and punctuation marks.

2.11.2 Tokenization

Tokenization divides the text into tokens, which might be individual sentences, words, phrases, or even characters. This process makes further analysis and feature extraction easier.

2.11.3 Stop word removal

Stop words are often used words such as ("the"," and", and "is") that have little significance in the context of the work at hand. Stop words are eliminated to concentrate on words that are more informative and to assist in minimizing the dimensionality of the data.

2.11.4 Lower casing

To guarantee that words with the same spelling but different capitalization are recognized as the same token, all text should be converted to lowercase. Consistency is improved and duplications are avoided as a result.

2.11.5 Stemming and lemmatization

Words are intended to be reduced to their root or basic form by lemmatization and stemming. While stemming uses heuristic principles to remove prefixes or suffixes, lemmatization creates legitimate words by taking the word's context and part of speech into account. These methods aid in reducing vocabulary size and normalizing word variances.

2.11.6 Removal of irrelevant parts

Depending on the demands of the assignment, it could occasionally be essential to delete particular text elements, such as URLs, numerals, or extraneous passages.

2.11.7 Handling acronyms and abbreviations

Within the document, having a large set of text may use acronyms or abbreviations that need to be clarified or changed to their full names to maintain consistency and clarity.

2.11.8 Spell-checking and correction

Spelling mistakes in text data may affect how well subsequent analysis works. These problems can be solved by utilizing spell-checking and other correction techniques.

2.11.9 Part-of-speech tagging

Giving terms in the text part-of-speech tags might provide more details about the sentence's grammatical structure. When performing more complex text analysis jobs, this might be helpful.

Depending on the kind of text data and the demands of the text summarizing task, several preparation techniques may be used. The quality of the output of the

summarization may be improved, and the performance of the following algorithms and models can be improved, by using the proper preprocessing procedures.

The following image shows the distribution of preprocessing used in text summarization.

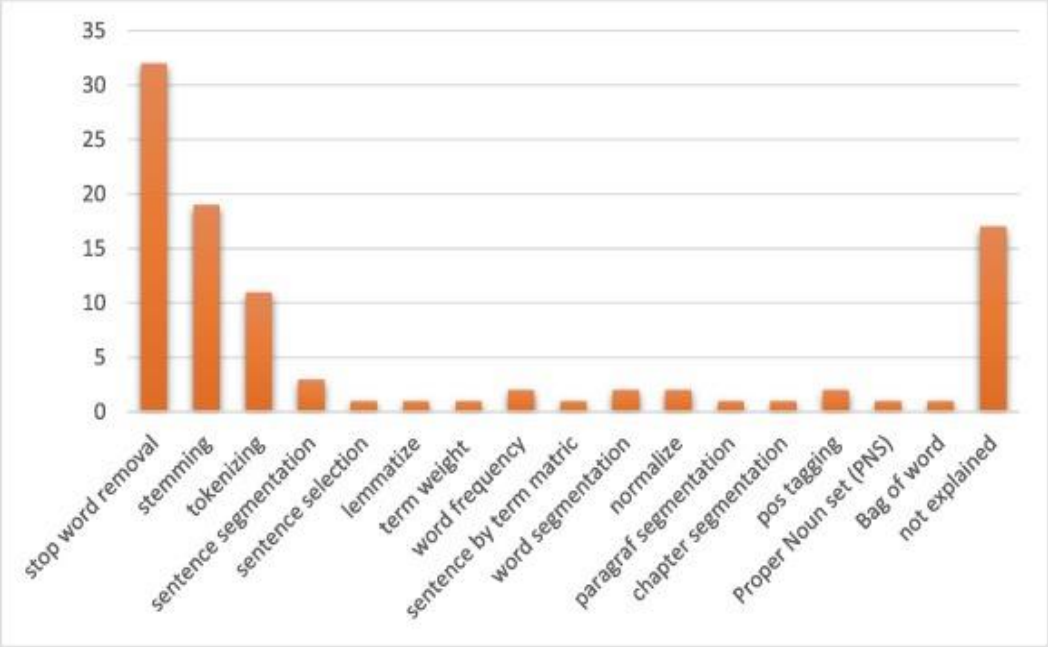


Figure 2: - Distribution of Preprocessing Used in Text Summarization (Widyassari et al., 2022).

2.12 Techniques for evaluation of text summarization

Text summarization, a crucial task in natural language processing (NLP), tries to reduce a lengthy text to a summary while preserving the essential details and core concepts. To gauge their performance and evaluate various methods, text summarizing systems must have high standards and be successful. The following are a few typical assessment strategies for text summarization:

2.12.1 Evaluation by sentence co-selection

The natural language processing (NLP) task of sentence co-selection entails choosing pertinent sentences from a given text or document based on a predetermined criterion. It is sometimes referred to as sentence extraction or sentence evaluation. Text summarization, information retrieval, and document analysis frequently employ this task.

Numerous criteria, including precision, recall, F1 score, and ROUGE scores, can be used in the evaluation.

- **Precision**

The percentage of chosen sentences that are pertinent or accurate serves as a measure of precision. It is obtained by dividing the total number of sentences chosen by the number of sentences that were correctly chosen.

- **Recall**

Recall gauges the percentage of pertinent sentences that are chosen. It is obtained by dividing the total number of pertinent sentences in the reference set by the number of correctly chosen sentences.

- **F1 score**

The F1 score, which gives a single metric to assess the system's performance, is the harmonic mean of accuracy and recall.

F1 score= 2 * (precision * recall) / (precision + recall).

- **ROUGE Score**

(Recall-Oriented Understudy for Gisting Evaluation) is a group of measures frequently used to rate text summarizing software. ROUGE-N calculates how many N-grams (contiguous sequences of N words, when N=1 Uni-gram, N=2 Bi-gram, and N=3 Tri-gram) are shared by the reference set and the sentences that were generated by the machine. ROUGE-L calculates the length of the common subsequence between the reference set and the sentences that were chosen by the machine.

During the assessment phase, the output of the system, which consists of the selected phrases, is compared with the reference set using the selected metrics. In terms of precision, recall, F1 score, and the overlap of N-grams or subsequences with the

reference set, the metrics offer numerical measurements of the system's performance. The machine performs better at choosing pertinent sentences as the scores rise.

2.12.2 Content-Based Evaluation

By using content-based similarity measures, we may address the co-selection measures' shortcomings that were previously highlighted. These techniques compare the two documents more closely than only by comparing their sentences. The fundamental approach is using the cosine similarity metric, which can be calculated using the following formula, to determine how similar the full-text content and its summary are.

$$\textbf{Cosine similarity (x, y)} = \frac{x.y}{||x||||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

2.13 Applications of text summarization

Text summarization has various application areas where it can be applied in the real world. To list some of them: -

1. Question and answering system (QA).
2. News summarization and news wire generation.
3. Report summarization for businessmen, politicians, and researchers.
4. Meeting summarization.
5. Automatic extraction and generation of titles.
6. Opinion summarization.

2.14 Related work

The development of an automatic text summarization model facilitated the work of humans as it shortens a long set of documents into short and precise contents that convey the whole document's information. Several types of research have been carried out to look at the summarization of text documents.

Extractive text summarization has been a popular topic of research in natural language processing, with a strong emphasis on global languages like English. Researchers have investigated a variety of methodological approaches, including statistical, rule-based, supervised machine learning techniques graph-based algorithms, and, more recently, deep learning-based models. These works have made significant contributions to the advancement of text summarizing techniques for major global languages. However, it remains relatively rare to research extractive summarization in local or under-resourced languages such as Awngi. It is unclear whether existing techniques and models can be applied to diverse local language contexts since many are designed and evaluated primarily for English or other well-resourced languages. The summarization of text in local languages can be challenging, as the datasets are usually smaller, the linguistic features are unique, and the cultural nuances are unique.

Text summarization technology must be adapted and evaluated in these contexts to ensure inclusivity and accessibility, especially in regions with diverse linguistic landscapes. There has been preliminary work on extractive summarization of local languages, including studies by (Demis, 2018) explored an automatic Amharic multi-document news text summarization using an open text summarizer. This study investigates the capability of the open-source tool (open text summarizer) for Amharic multi-document summarization to fill in the gaps in single-document text summarization. There were 35 Amharic single-text news articles collected from Ethiopian news reporters, Addis Admass, Addis Zena, and Walta information centers for this study. The researcher re-coded the system and redesigned the user interface of the open text summarizer using the C# programming language and Visual Studio 2010. A sentence extraction-based summarization was introduced for Amharic news which uses open-text summarization. Intrinsic evaluation techniques were used to assess the performance of the customized open-text summarizer. In general, the objective and

subjective assessment metrics from the summarizer have yielded positive results for the researchers.

Another study (Redi, 2020) demonstrated extractive Amharic text summarization using latent Dirichlet allocation. To extractive summarization of Amharic text, this work presents a probabilistic topic-modeling approach called Latent Dirichlet Allocation (LDA). For investigation, the study utilized 60 news documents collected from different newspapers. The Python programming language was employed for coding purposes. The system was evaluated using the F-measure, an objective assessment tool. In conclusion, the study suggested the proposed approaches perform better than the earlier research due to the integrated features and varied strategies.

A published journal article by (Guadie et al., 2021) manifested Amharic text summarization for news items posted on social media. The summarization technique was applied as an extractive summarization approach, which assigns sentences from the posted documents with the highest ranking to form summaries, and the user can identify the size of the summary. Twitter and Facebook were the main sources of documents. A total of 4951 short sentences were collected in different news items from Twitter and Facebook, covering protests (3943), droughts (667), floods (101), and sports (246). To evaluate the performance of the system, the researchers used subjective (qualitative) and objective (quantitative) methods. Informational quality, coherence of structure, and linguistic quality of the automatic summary were measured via subjective evaluations. In contrast, objective evaluations measure the performance of the standard for precision and recall in the given input texts. The assessment method has demonstrated that this yields excellent results for summarizing texts that have been uploaded on social media.

Afaan Oromo Text Summarization using Word Embedding, a thesis by (Tashoma et al., 2020b). It is a generic automatic text summarizer based on the word-to-vec model. To test

the performance of the model both primary and secondary sources of data were utilized. The researchers collected 22 different documents from the internet and newspapers, among those 13 were used as validation and 9 were used as testing. The model was evaluated using both subjective and objective summary evaluation metrics. By employing the same data as the earlier studies, the summarizer outscored them overall in terms of precision, recall, and F-measure by 0.648, 0.626, and 0.058, respectively. Even though a comparison was made with previous studies, there was no clear evidence, and a benchmark study was discussed.

As a result of these efforts, certain techniques have proven viable when applied to local language datasets, however, further research and innovation are necessary to address the unique challenges presented by these languages. Toward filling this gap, the current study develops and evaluates a prototype system that can effectively handle local language specifically for Awngi text documents as part of an effort to broaden the scope of extractive text summarization research. The evaluation of the proposed prototype system involves testing the performance of machine-generated summary with human reference summary based on rouge and cosine similarity which previous researchers failed to fill.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Introduction

Regardless of what type of scientific investigation is being conducted, research methodology is crucial to ensuring a well-designed, reliable, and valid study. Hereafter, the precise steps or methods used to find, choose, process, and evaluate data on a subject are discussed. In addition to this, the kind of research design methodology, where and how the data was organized and pre-processed, the tools and techniques that have been used, and the metrics used to measure the performance of the model are illustrated.

3.2 Literature review

To have a conceptual understanding of the types of automatic text summarization approaches, preprocessing, model development, and evaluation of the model, we have covered different materials, including journals, articles, conference papers, newspapers, and more resources. In addition to this, the state of the art in automatic text summarization was assessed. Moreover, to understand the nature of the Awngi language, the alphabets used, and sentence formation books written by different authors, and journals were reviewed.

3.3 System architecture

It is the overall design and structure of a complex system that includes the arrangement of its components, subsystems, modules, and interfaces, as well as the interrelationships among those components. The system architecture illustrates how its components work together, how they perform the desired function, and how they interact with each other. The entire architecture for the proposed system is depicted in the following figure.

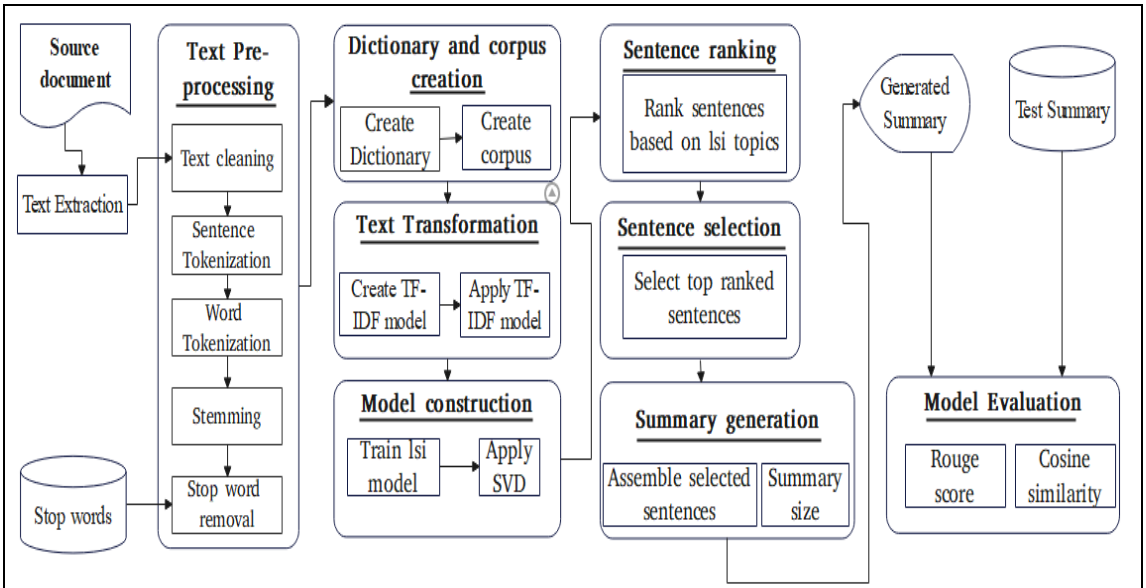


Figure 3: - System Architecture

3.4 Research design methodology

For this study, design science research design methodology was conducted. a design science research methodology was applied since, within this methodology, a problem was assessed, and after the identification of a certain problem, an artifact was developed so that the artifact would be evaluated. Tools, techniques, approaches, algorithms, and assessment processes are all part of design science.

3.5 Dataset collection and preparation

The source of data for this research work was collected from the publicly known website Amhara Media Corporation (አሚኮ), which contains a webpage called Cherbewa (ቸርቤዋ አዋኛ). The second source of documents was a printed magazine. The hard copy was collected from a magazine shop and library and captured with a camera. After that, the text from the image file was extracted using tools like Google Sheets and the MetaAppz Amharic OCR website.

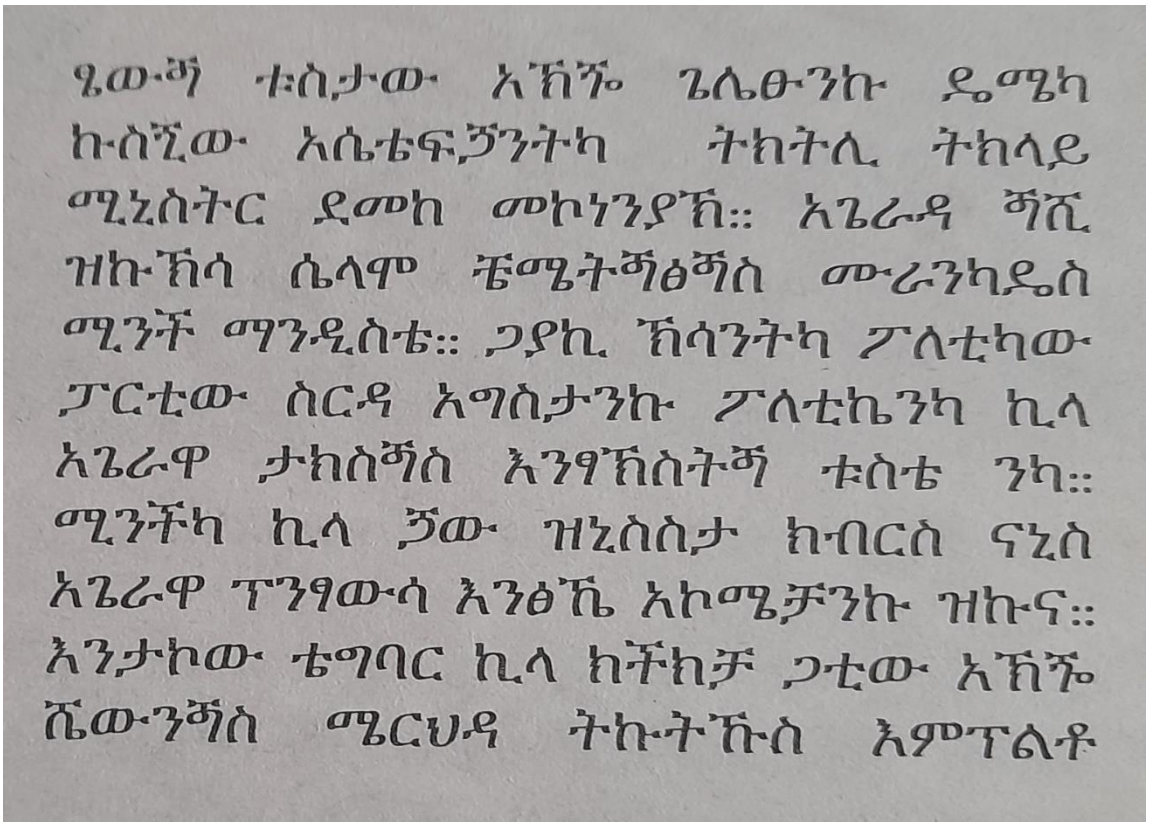
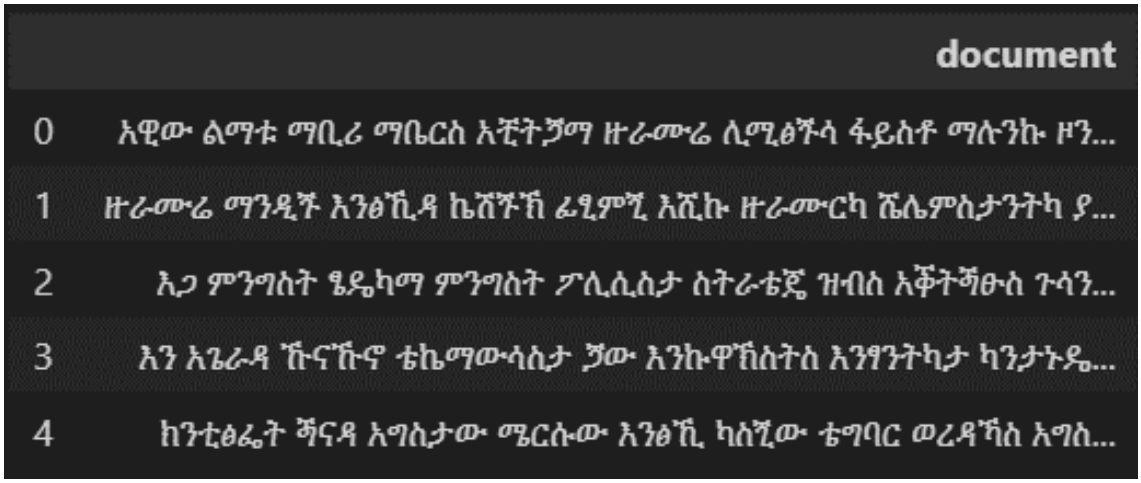

ዒውሓ ቱስታው አኽኼ ጌሌፀንኩ ዴሜካ
ኩስኺው አሴቴፍቻንትካ ትክትሊ ትክላይ
ሚኒስትር ደመካ መኮነንያኽ። አጌራዳ ሻሺ
ዝኩኽሳ ሴላሞ ቼሜትሻፅሻስ ሙራንካዴስ
ሚንች ማንዲስቴ። ጋያኪ ኽሳንትካ ፖለቲካው
ፓርቲው ስርዳ አግስታንኩ ፖለቲኬንካ ኪላ
አጌራዋ ታክስሻስ እንፃኽስትሻ ቱስቴ ንካ።
ሚንችካ ኪላ ቻው ዝኒስስታ ክብርስ ናኒስ
አጌራዋ ፕንፃውሳ እንፅኼ አኮሜቻንኩ ዝኩና።
እንታኮው ቴግባር ኪላ ክችክቻ ጋቲው አኽኼ
ሼውንሻስ ሜርህዳ ትኩትኹስ እምፕልቶ

Figure 4: - Sample image data

To make the collected documents ready for analysis the documents are organized in a (CSV) format. Figure 5 below demonstrates multiple documents in a CSV file format, this enables us to easily analyze using pandas data frame.



	document
0	አዊው ልማቱ ማቢሪ ማቤርስ አቺትጃማ ዙራሙሪ ሊሚፅኾሳ ፋይስቶ ማሉንኩ ዞን...
1	ዙራሙሪ ማንዲኾ እንፅኺዳ ኬሸኾኽ ፊሚሚ እሺኩ ዙራሙርካ ሼሌምስታንትካ ያ...
2	እጋ ምንግስት ፄዴካማ ምንግስት ፖሊሲስታ ስትራቴጂ ዝብስ አቕትሻፁስ ጉሳን...
3	እን አጌራዳ ኹናኹኖ ቴኬማውሳስታ ጋው እንኩዋኽስትስ እንፃንትካታ ካንታኑዴ...
4	ከንቲፅፌት ሻናዳ አግስታው ሜርሱው እንፅኺ ካስሚው ቴግባር ወረዳኻስ አግስ...

Figure 5: - Document in CSV format

3.5.1 Manual summary preparation

As opposed to automatic summaries generated by computers or software, manual summaries are created by humans. After reading and comprehending the original text, an individual condenses the main points, key ideas, and relevant information into a concise and coherent summary. A manual summary captures the essence of the original text in a condensed way to convey its key information effectively.

Extracting and presenting the most important parts of a text requires human judgment, understanding, and language skills. In various contexts, such as journalism, academic writing, or content curation, manual summaries are frequently used as references to evaluate the performance of text summarization algorithms or as standalone summaries. To perform a human reference summary, experts with the language were involved. Based on the guidelines provided in the appendices B, the manual summary was taken with a compression rate of 20%, 30%, and 40%. For manual summary preparation, the guidelines were adopted from (Computing & Guadie, 2017; Demis, 2018; Guadie et al., 2021; Yirdaw & Ejigu, 2012).

3.5.2 Data preprocessing steps

Prior to usage, data preparation is necessary. The idea of data preprocessing is to transform unprocessed data into clean data collection. Before the dataset is sent to the algorithm, it is preprocessed to look for missing values, noisy data, and other irregularities. All data must be in machine-appropriate formats.

- **Sentence tokenization**

  Sentence tokenization is the process of splitting a set of documents or paragraphs into sentences. The sentence delimiters we have used are (? ። !). Moreover, (።) aratnetib exists in different forms such as (::) two colons together, (: :) two colons separated by space, and a single standalone aratnetib (።). So, using regular expressions we have replaced those different forms of aratnetib into a single standalone aratnetib. Since NLTK (Natural Language Tool Kit) has no built-in module for sentence tokenization for Awngi text documents, we have used a customized PunktSentenceTokenizer from NLTK.

  Example: - The following is a sample paragraph having four sentences delaminated by (። and ?).

  para='እን ዴብቱርት እምፐል ያኽ። እን ዴብቱርት ላኝ ያኽ። እን ዴብቱርት ሹካኽ። እንኪ እሊዉ ዴብቱርት ያኽ?'

  After a sentence tokenization is applied it will have a list of four sentences.

```
['እን ዴብቱርት እምፐል ያኽ።',
 'እን ዴብቱርት ላኝ ያኽ።',
 'እን ዴብቱርት ሹካኽ።',
 'እንኪ እሊዉ ዴብቱርት ያኽ?']
```

- **Word tokenization**

  Word tokenization involves splitting a large piece of text into words, which is required for natural language processing tasks like classifying and counting words. In this

research, we have utilized NLTK's TweetTokenizer since it captures punctuations that are not part of English punctuation marks such as (። ፤ ፥).

Example: - The above sentences can be further tokenized into words as follows.

[['እን', 'ዴብቴርት', 'እምፐል', 'ያኽ', '።'],
['እን', 'ዴብቴርት', 'ላኟ', 'ያኽ', '።'],
['እን', 'ዴብቴርት', 'እምፐል', 'ሹኻኽ', '።'],
['እንኪ', 'እሊዉ', 'ዴብቴርት', 'ያኽ', '?']]

❖ **Acronym expansion**

Acronyms are abbreviations created by combining the first letter(s) of one or more words to generate a new term. Instead of pronouncing the resultant word as separate letters, it is spoken as a whole. Acronyms are frequently used to condense and simplify larger words or sentences. Preprocessing and removing acronyms is important, as the symbols contained within acronyms are not necessary for the meaning. By removing these symbols, the remaining text becomes meaningless without expanding the acronyms. Therefore, preprocessing and removing acronyms should be a standard step to ensure the text maintains its intended meaning. To expand acronyms properly, it is necessary to use a dictionary or list that maps each acronym to its corresponding expanded words. By referencing this dictionary during preprocessing, the acronyms can be replaced with their full expansions. This ensures the text maintains its intended meaning after removing the unnecessary symbols contained within the original acronyms. The following table depicts sample acronyms in the Awngi writing system.

Table 9: - Awngi acronyms and its expansions

Acronym	Expansion
ም/አ	ምሬት አሜት
ፅ/ኻና	ፅፈት ኻና
ዊ/ሩ	ዊዛሩ
ዶ/ር	ዶክተር
ፍ/ኻና	ፍርድ ኻና

❖ **Normalization**

As a way to ensure consistency and reduce variations between words or phrases, text normalization involves applying several techniques to transform data into a standardized or canonical format. Even though the Awngi alphabet does not contain such variants, while writing, such a kind of mixing appears. The alphabet and its variants are discussed in the table below.

Table 10: - Alphabets and its variants

Alphabet	Variants	Example word	Normalized
ሀ	ሀ፣ ሐ ፣ ኀ	ሀገረስብኬት ፣ ሐገረስብኬት ፣ ኀገረስብኬት (ሀገረስብከት)	ሀገረስብኬት
ሰ	ሰ ፣ ሠ	ሴኒ ፣ ሤኒ (ሰኔ)	ሴኒ
አ	አ ፣ ዐ	አይሊኒ ፣ ዐይሊኒ (ሀይለኛ)	አይሊኒ
ፀ	ፀ ፣ ጸ	ፅራግ ፣ ጽራግ (መጥረጊያ)	ፅራግ

❖ **Text cleaning**

Text **c**leaning involves removing unnecessary punctuation marks, digits, and short sentences that don't add any value to the model development. The punctuations such as (!"#$%&'() *+, -. /: ;<=>? @ [\] ^_` {|} ~ ፡ ። ፤ ፥ ፦ ፧ ※) are removed. To remove punctuations, we have used Python's regular expression and string module.

Furthermore, we got alphabets that were wrongly written, such as the alphabet 'ዐ' was interpreted as the numeric '0' (zero). for example, to write 1000 it was written as 1ዐዐዐ. 'ሻ' as 'ዅ', for instance, the word 'አክዅ' to write as 'አክሼ' and 'ኌ' as ('ሻ|ዅ|ኌ') e.g., 'አክኌ/አክዅ/አክሻ'. Those errors exist because of an OCR application that we have used while transforming a hard copy into a soft copy.

❖ **Document extraction**

Since the research is focused on summarizing long documents into shorter versions, it was necessary to remove any documents that were too short. To accomplish this, an algorithm was used to identify and remove any documents that fell below a certain length threshold. This ensured the dataset only contained longer documents that could be effectively summarized, without the noise of very short documents that would not provide enough information for the summarization task. The following Figure 6 depicts this an algorithm for removing too short documents.

Algorithm 1: Algorithm for removing too short documents

Input: *a dataframe column having a list of sentences*

Output: *a dataframe column of documents having the minimum number from the mean.*

#sen_column ← calculate total number of sen per doc

average_num_of_sen ← calculate mean

If *#sen_column* ***less than*** *average_num_of_sen*

Pass

else

Include with in the document list

i.e, each variable is a list of values

Figure 6: - An algorithm for removing too short documents

❖ **Sentence extraction**

As we know words are the building blocks for sentences. So, longer sentences are those that have too many words in them whereas short sentences are those that have too few words. For the sake of extracting sentences that are reliable for summarization, we have developed an algorithm that removes sentences with too many words and sentences with too few words.

Algorithm 2: -Algorithm to remove too long and too short sentences from doc.

Input: *a dataframe column having a list of sentences*

Output: *list of sentences*

```
For word in sentence.split()
    #words_in_sen  ← len(word)
     max_words_per_sen  ← Max(#words_in_sen)
     min_words_per_sen  ← Min((#words_in_sen)

average_of_max_words_per_sen  ←  mean(max_words_per_sen)
average_of_min_words_per_sen  ← mean( min_words_per_sen)

selected_sentences  ←  [word For word in sentence if
len(word.split())>=average_of_min_words_per_sen and
len(word.split())<=average_of_max_words_per_sen]
```

Figure 7: - An algorithm for removing too-long and too-short sentences

❖ **Stop word removal**

The stop words in any language, not only English are a collection of commonly used words such as (for, an, nor, but, or, yet, and so). As compared to English Awngi text documents also have a stop-word list. Since there was no pre-existing list of stop words available, a custom set of stop words was prepared for this research. To generate the stop word list, the total frequency of all words within the documents was calculated. The 47 words with the highest overall frequency were then selected and designated as the stop words to be removed during preprocessing. This custom stop word list ensured the most common and least informative words were filtered out, allowing the summarization

algorithm to focus on the most relevant content within the long documents. The following table contains a sample list of stop words in Awngi.

Table 11: - Sample Awngi stop word list

ምክናያትኪ	አኒ	አን
አንዳ	አኹኪ	አኽሻስ
አኽኹ	እስታ	እኒ
እን	እንት	እንዶጂሱ

3.6 Stemming

As part of natural language processing (NLP) and information retrieval, stemming reduces words to their root or basic form, known as a "stem." The stem does not contain a complete word by itself, but it contains the core meaning of the word. It enables words to be normalized by recognizing variations in the same word, regardless of inflections or suffixes, as the same word. The text data is less dimensionalized as a result, and text analysis tasks like retrieval, search engine, classification, and sentiment analysis are improved. In the example of "running," "runner," and "runs," stemming would reduce these words to their common stem "run." Similarly, words such as "jumping," "jumps," and "jumped" will all stem to "jump. a rule-based stemmer was developed and used to remove suffixes from words. This stemming process helped normalize the vocabulary by reducing words to their base or root forms. This was an important preprocessing step to ensure semantically similar words were treated equivalently by the summarization model, regardless of their specific grammatical forms.

Table 12: - Sample Awngi stemming for removing suffixes

Word	Root word	Suffixes
ይኮኖሚው፣ ይኮኖሚውሳ፣ ይኮኖሚዳ፣ ይኮኖሚስ፣ ይኮኖሚያዊ፣ ይኮኖሚያቄ፣ ይኮኖሚውዳ፣ ይኮኖሚክሱ	ይኮኖሚ	["ው","ውሳ","ዳ","ስ","ያዊ","ያቄ","ውዳ","ክሱ"]
ጉሳንቲ፣ ጉሳንቱሳ፣ ጉሳንኩሳ፣ ጉሳንኩ፣ ጉሳንታ፣ ጉሳንትካ፣ ጉሳንቲዬስ	ጉሳን	["ቲ","ቱሳ","ኩሳ","ኩ","ታ","ትካ","ቲዬስ"]

3.7 Feature extraction methods

Involves selecting and transforming raw data into meaningful and representative features. Feature extraction plays an integral role in analyzing and understanding data in a variety of fields, including machine learning, computer vision, and natural language processing. Text analysis comprises several feature extraction methods. Even though we employed TF-IDF feature extraction technique to extract the salient features, here are a few popular commonly used techniques: -

- **Bag-of-words: -**

  The Bag of Words (BoW) approach is widely used in natural language processing (NLP) and information retrieval. It is a straightforward and efficient method for encoding textual data as a numerical feature vector. It treats a text as an unordered collection or "bag" of words, ignoring grammar and word order and focusing solely on word frequency. In this paradigm, a document is represented as a vector, with each element representing a unique word from the text's vocabulary. The value of each element in the vector indicates the frequency or existence of that term in the document.

Text	ላጃንቲ	እምፐላንቲ	እን	ከላ	ዝኮ	ያኽ	ዴብቴርት
እን እምፐላንቲ ዴብቴርት ያኽ።	0	1	1	0	0	1	1
ላጃንቲ ዴብቴርት ከላ ዝኮ።	1	0	0	1	1	0	1

Figure 8: - Feature extraction using BOWs

- **Term Frequency-Inverse Document Frequency (TF-IDF)**

It is a numerical statistic used in natural language processing (NLP) to evaluate the importance of a word within a collection or corpus of documents. The TF-IDF incorporates both a term's frequency in a document (TF) as well as its rarity across the entire corpus (IDF). Even though both Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction methods were considered, the research utilized TF-IDF feature extraction.

TF-IDF = (Term Frequency in Document) * (Inverse Document Frequency)

Text	ላሻንቲ	አምፐላንቲ	አን	ከለ	ዝኮ	ያሽ	ዴብቴርት
አን አምፐላንቲ ዴብቴርት ያሽ።	0	0.534046329	0.534046329	0	0	0.534046329	0.379978362
ላሻንቲ ዴብቴርት ከለ ዝኮ።	0.534046329	0	0	0.534046329	0.534046329	0	0.379978362

Figure 9: - Feature extraction using TF-IDF

3.8 Model development

A model development process involves designing, creating, and refining a model for a specific task or problem. It involves several steps and considerations to create a precise and effective model.

Genism's LSI (Latent Semantic Indexing) model is a method used to analyze textual data in terms of semantics and dimension reduction. Additionally, it is also known as Latent Semantic Analysis (LSA). LSI is a method for retrieving information and comparing documents that are based on the concept of singular value decomposition (SVD). The LSI model in genism involves the following steps: -

- **Corpus creation: -**

  First, you must establish a corpus, which is a collection of text documents.

- **Document-Term Matrix: -**

A document-term matrix is then produced, in which each row represents a document, and each column represents a term within the corpus. The values represent the term frequency or some weighted representation of the terms.

- **Singular Value Decomposition (SVD): -**

In the LSI model, singular value decomposition is applied to the document-term matrix. Singular value decomposition decomposes the matrix into three separate matrices: U, S, and V. U is the matrix of the document topic, S is the matrix of the singular values, and V is the matrix of the term topic.

- **Dimensionality Reduction: -**

To capture the latent semantics of each document, LSI selects a subset of singular values from the U and V matrices and their corresponding columns.

- **Semantic Analysis: -**

A semantic analysis of documents and terms is performed on reduced-dimensional matrices derived from SVD. The LSI model represents documents and terms as dense vectors in the reduced-dimensional space.

- **Similarity Calculation: -**

LSI may be used to calculate document or phrase similarity using the cosine similarity measure. The cosine similarity between two document or phrase vectors can be used to identify their semantic similarity or relatedness.

3.9 Tools used

To preserve this study Microsoft Word 2019 was used for documentation, Python 3.9 as a programming language, and Visual Studio code editor. Why Python 3.9 is that the Python package for NLP which is NLTK, works with Python 3.9 and lower versions and Visual Studio code editor since it has an interactive interface and all necessary Python extensions. Pandas data frame which is a two-dimensional data structure was used for the analysis

purpose. NLTK (Natural Language Toolkit) facilitates working with human language data in Python. Tkinter (tkinter) is Python's standard GUI (Graphical User Interface) package. It is a built-in module included in the Python standard library. Tkinter offers a variety of widgets and tools for developing desktop apps with graphical interfaces. Hardware tools that were used for this study include HP Laptop Core i5, with 238 Gb SSD included for fast performance. For visualization purposes, the Matplotlib library is used having static, animated, and interactive graphics. It enables both difficult and easy tasks.

3.10 Techniques

Using text summarization, you can condense a given text without losing its key information and meaning.

3.10.1 LSA (Latent Semantic Analysis)

It is a natural language processing technique that examines connections between a group of texts and the terms used in them. It scans unstructured material using the mathematical approach of singular value decomposition to look for undiscovered connections between phrases and ideas (*Latent Semantic Analysis (LSA)*, 2023). LSA is named after the fact that SVD, when applied to document-word matrices, groups documents that are semantically related to each other, even if they don't have a common word (V. Gupta & Lehal, 2010). “LSA is an unsupervised technique that represents text semantics based on the observed co-occurrence of words”(El-Kassas et al., 2021). This study uses the LSA (Latent Semantic Analysis) method that analyzes relationships between documents and the terms contained within them using natural language processing. A mathematical technique called singular value decomposition is used to scan unstructured data for hidden relationships.

Utilizing LSA vectors for summarization has the advantage over using word vectors because the conceptual (or semantic) relations represented in the human brain are

automatically captured in the LSA, whereas using word vectors without the LSA transformation necessitates the creation of explicit methods to derive conceptual relations.

3.11 Performance metrics

The text summary is the process of reducing the length of a given text while maintaining the major ideas and important details. The effectiveness of text summarizing systems must be assessed using particular measures. To measure the performance of the summary we have used two metrics.

3.11.1 ROUGE

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is a collection of measures that are frequently used to assess the quality of automatic summaries by contrasting them with reference summaries created by humans. It calculates the amount of overlap between the generated summary and the reference summary's n-grams (contiguous sequences of n words). Popular variations of ROUGE-1, ROUGE-2, and ROUGE-L take into account bigrams, unigrams, and longest common subsequences, respectively.

Example: - let’s assume the following machine, and human-generated summaries as,

System summary (ss)= ['አበበ', 'ዳታ', 'ሳይንሶ', 'እንካኔ']

Human summary (hs)= ['አበበ', 'ዳታ', 'ሳይንሶ', 'ኢይሎ', 'እንካኔ']

Overlapping words (ow) = ['አበበ', 'ዳታ', 'ሳይንሶ', 'እንካኔ']

Total number of overlapping words (#ow) = 4

Total number of words in system summary (#wss) = 4

Total number of words in human summary (#whs) = 5

ROUGE-1 (unigram):

ROUGE-1 calculates the overlap of unigrams (single words) between a candidate summary and a reference summary.

$$\text{Recall} = \frac{\text{Total number of overlapping words (\#ow)}}{\text{Total number of words in human summary (\#whs)}} = \frac{4}{5} = 0.8$$

$$\text{Precision} = \frac{\text{Total number of overlapping words (\#ow)}}{\text{Total number of words in system summary (\#wss)}} \quad \frac{4}{4} = 1$$

$$\text{F1 score} = \frac{2*(\text{precision}*\text{recall})}{(\text{precision}+\text{recall})} \quad \frac{2(1*.8)}{1+.8} = 0.88$$

ROUGE-2 (bigram):

As opposed to ROUGE-1, ROUGE-2 calculates the overlap of bigrams (two words) between a candidate summary and a reference summary.

System summary (ss)= ['አበበ ዳታ', 'ዳታ ሳይንሶ', 'ሳይንሶ እንካኔ']

Human summary (hs)= ['አበበ ዳታ', 'ዳታ ሳይንሶ', 'ሳይንሶ አይሎ', 'አይሎ እንካኔ']

Overlapping words (ow) = ['አበበ ዳታ', 'ዳታ ሳይንሶ']

Total number of overlapping bigrams (#ow) = 2

Total number of bigrams in system summary (#wss) = 3

Total number of bigrams in human summary (#whs) = 4

$$\text{Recall} = \frac{\text{Total number of overlapping bigrams (\#ow)}}{\text{Total number of bigrams in human summary (\#whs)}} \quad \frac{2}{4} = 0.5$$

$$\text{Precisio} = \frac{\text{Total number of overlapping bigrams (\#ow)}}{\text{Total number of bigrams in system summary (\#wss)}} \quad \frac{2}{3} = 0.66$$

$$\text{F1 score} = \frac{2*(\text{precision}*\text{recall})}{(\text{precision}+\text{recall})} = \frac{2(0.66*0.5)}{0.66+0.5} = 0.56$$

Rouge-L

As its name implies, ROUGE-L finds the longest common subsequence (LCS) between the output of our model and the reference, that is, the longest sequence of words (not necessarily consecutive, but still in order) that are shared by both. A longer shared sequence likely indicates a greater degree of similarity.

3.11.2 Cosine similarity

A typical metric for determining how similar two vectors are in a multi-dimensional space is called cosine similarity. Natural language processing (NLP) activities like text summarization and document retrieval frequently use it.

Cosine similarity can be used to evaluate the semantic similarity between the generated summary and the reference summaries in the context of text summarization. The following steps show how to perform a cosine similarity between two documents.

Step 1: - Transform the given document into vectors.

Example: - docs = ['አበበ ዳታ ሳይንሶ እንካኔ።','አበበ ዳታ ሳይንሶ አይሎ እንካኔ።']

*Table **13**: - Features and their vectors using TF-IDF*

	ሳይንሶ	አበበ	አይሎ	እንካኔ	ዳታ
0	0.50000	0.50000	0.000000	0.50000	0.50000
1	0.40909	0.40909	0.574962	0.40909	0.40909

Now, we can calculate the cosine similarity between each sentence within the document using the cosine formula.

Cosine similarity (x, y) = $\frac{x.y}{||x||||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$

Even though the above formula can perform the cosine similarity, we have used the SKLEARN'S built-in method. So, the following table illustrates the similarity among the sentences within the document.

*Table **14**: - Cosine similarity between sentences*

	አበበ ዳታ ሳይንሶ እንካኔ።	አበበ ዳታ ሳይንሶ አይሎ እንካኔ።
አበበ ዳታ ሳይንሶ እንካኔ።	100.0	81.0
አበበ ዳታ ሳይንሶ አይሎ እንካኔ።	81.0	100.0

CHAPTER FOUR

4. RESULT AND DISCUSSION

4.1 Introduction

By selecting and extracting the most salient sentences from the original text, extractive text summarization is used to automatically generate a concise summary of a document. Natural language processing faces this fundamental problem in several applications, such as summarizing news articles, academic papers, and customer reviews. Although significant progress has been made in extractive summarization, the current state-of-the-art models continue to struggle to consistently produce coherent, high-quality summaries, especially for longer documents. It remains a priority to research how extractive summarization systems can be improved in terms of performance and robustness. In this research, we design and implement an extractive text summarization system for Awngi news documents.

This section presents the results of our experiments, demonstrating how our extraction summarization model outperforms other approaches. We respond to each of the study questions and offer an in-depth analysis of the efficacy of various strategies and methods used in the summarizing procedure.

4.2 The data source for Awngi text documents

To create a representative dataset for Awngi text summarization, we have collected Awngi text documents from various sources, ensuring that the dataset covers different domains. Amhara Media Corporation's (አማኮ) website is the main source; additionally, hard copies of newspapers were collected from libraries and shops. A total of 558 documents were collected from the above sources, and through preprocessing and removing lengthy and too-short sentences, the resulting dataset contains a total of 213 documents.

4.2.1 Document Statistics

The document statistics in Table 15 below show the overall data that was collected from cherbewa magazine. The dataset contains 213 documents with a wide range in the number of sentences and words. On average, each document has 45 sentences and 760 words, with an average sentence length of 18 words. Understanding these characteristics can help us understand how this dataset might be used for different text analysis tasks.

Table 15: - Source document statistics

Document Attributes	Value
Number of documents	213
Maximum number of sentences per document	85
Minimum number of sentences per document	30
Mean of sentences per document	45
Maximum number of words per document	1301
Minimum number of words per document	353
Mean of words per document	760
Maximum number of words per sentence	44
Minimum number of words per sentence	5
Mean of words per sentence	18

4.2.2 Human reference summary statistics at (20%) extraction rate

Table 16 below demonstrates human reference summary statistics at a 20% extraction rate. Among the 213 documents in this dataset, there are 6-17 sentences and 81-264 words. The average document consists of 9.399 sentences and 155.69 words, with an average of 18 words per sentence.

Table 16: - Human reference summary statistics at (20%)

Document Attributes	Value
Number of documents	213
Maximum number of sentences per document	17

Document Attributes	Value
Minimum number of sentences per document	6
Mean of sentences per document	9.399
Maximum number of words per document	264
Minimum number of words per document	81
Mean of words per document	155.69
Maximum number of words per sentence	44
Minimum number of words per sentence	5
Mean of words per sentence	18

4.2.3 Human reference summary statistics at (30%) extraction rate

From the overall 213 documents, Table 17 shows the number of human reference sentences ranging from 9 to 26 and 109 to 416 words, with an average of 13.939 sentences and 233.667 words per document at an extraction rate of 30%.

Table 17: - Human reference summary statistics at (30%)

Document Attributes	Value
Number of documents	213
Maximum number of sentences per document	26
Minimum number of sentences per document	9
Mean of sentences per document	13.939
Maximum number of words per document	416
Minimum number of words per document	109
Mean of words per document	233.667
Maximum number of words per sentence	44
Minimum number of words per sentence	5

Document Attributes	Value
Mean of words per sentence	18

4.2.4 Human reference summary statistics at (40%) extraction rate

In the same way, as the above two extraction rates (20% and 30%) Table 18 also shows among 213 total documents the average number of sentences is 18.357 per document with an average 307.883 number of words per document with an extraction rate of 40%.

Table 18: - Human reference summary statistics at (40%)

Document Attributes	Value
Number of documents	213
Maximum number of sentences per document	34
Minimum number of sentences per document	12
Mean of sentences per document	18.357
Maximum number of words per document	540
Minimum number of words per document	157
Mean of words per document	307.883
Maximum number of words per sentence	44
Minimum number of words per sentence	5
Mean of words per sentence	18

4.2.5 LSA Summary statistics at 20% extraction rate

Table 19: - LSA Summary statistics at 20% extraction rate

Document Summary Attributes	Value
Number of summary document	213
Maximum number of sentences per summary	17
Minimum number of sentences per summary	6

Document Summary Attributes	Value
The Mean of sentences per summary	9.399
Maximum number of words per summary	410
Minimum number of words per summary	109
The mean of words per summary	250.582
Maximum number of words per sentence	44
Minimum number of words per sentence	23
The Mean of words per sentence	37.46

4.2.6 LSA Summary statistics at 30% extraction rate

Table 20: - LSA Summary statistics at 30% extraction rate

Document Summary Attributes	Value
Number of summary document	213
Maximum number of sentences per summary	26
Minimum number of sentences per summary	9
The Mean of sentences per summary	13.93
Maximum number of words per summary	564
Minimum number of words per summary	134
The mean of words per summary	338.25
Maximum number of words per sentence	44
Minimum number of words per sentence	5
The Mean of words per sentence	16.45

4.2.7 LSA Summary statistics at 40% extraction rate

Table 21: - LSA summary statistics at 40% extraction rate

Document Summary Attributes	Value
Number of summary document	213
Maximum number of sentences per summary	34
Minimum number of sentences per summary	12
The Mean of sentences per summary	18.35
Maximum number of words per summary	670
Minimum number of words per summary	153
The mean of words per summary	410.51
Maximum number of words per sentence	44
Minimum number of words per sentence	5
The Mean of words per sentence	16.45

4.3 Experimental setup

In the rapidly evolving digital landscape, the integration of advanced hardware and software technologies has opened up new frontiers for scientific research and data-driven discovery. At the core of our investigation was a high-performance computing system, comprising an HP 250 G7 Notebook equipped with an Intel(R) Core (TM) i5 processor, 12GB of DDR4 RAM, a 238 GB solid-state drive, and 1TB of HDD drive running on windows 10 pro-64-bit operating system. As a result of the computational capabilities of this platform, the study was able to handle the complex data processing and analysis.

4.4 Human reference summary document preparation system

To compare the output of automated text summarization systems with that of human reference summaries, we can refer to the human reference summary. They are used as a

benchmark for measuring the performance of the system-generated summary. For the sake of easy, non-redundant, and error-free reference summary preparation we have developed a system. Figure 10 below shows a desktop application system developed using the Python programming language specifically using the Tkinter module. To generate a human (reference) three Awngi language experts were selected. As we have a total of 213 documents each expert was assigned 71 documents. Having multiple experts makes it comprehensive. while doing so the expert was given the source documents via this system. Once given, he/she loads each document, and the documents are displayed as a sentence in parallel with a checkbox to select or deselect the given sentence. After that selecting the extraction rate (20%,30%, or 40%) is necessary from the dropdown menu. The figure below demonstrates the overall system user interface and sample loaded document for summary.

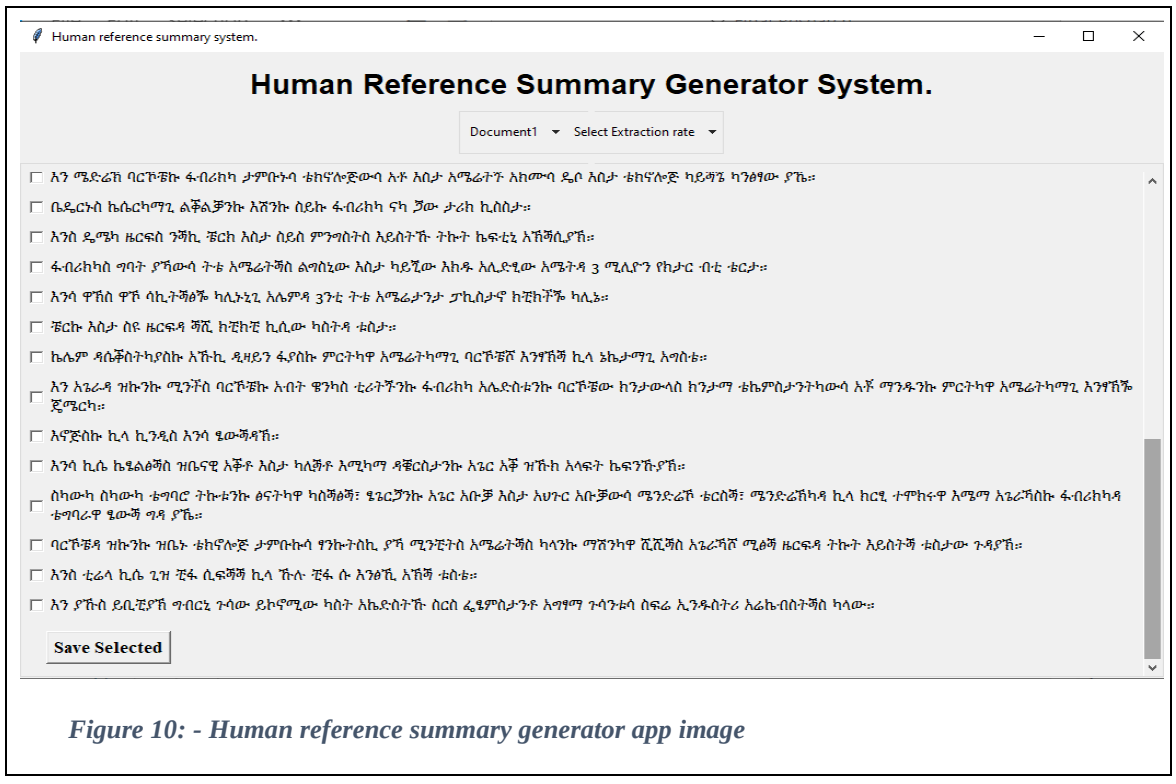


Figure 10: - Human reference summary generator app image

One of the human reference summary generator system's qualities is that, it automatically disables the checkboxes while reaching the maximum extraction rate (20%,30%, or 40%). The following Figure 11 depicts a human reference summary at a 20% summarization rate.

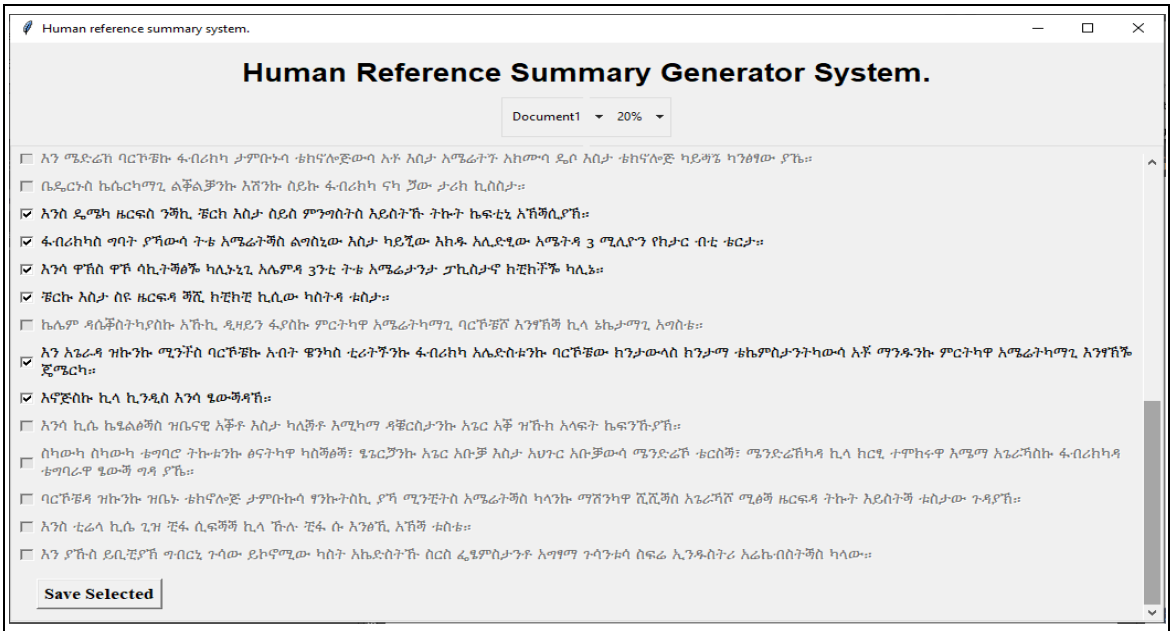


Figure 11: - Human reference summary generation at 20% image

After a while, he/she should have to save the selected sentence for each extraction rate. For simplicity, the system creates a folder for each document at the extraction rate of (20%, 30%, and 40%).

4.5 Rouge Performance evaluation

A summary of a document is a concise paragraph that highlights key information from the original content (M. Gupta & Garg, 2016b). An extractive text summarization extracts key sentences or phrases from a text and retains only the most important content. To do so there are commonly used extractive text summarization techniques. *Frequency-based (TF-IDF) technique,* according to the technique, sentences are scored based on the frequency of significant words across documents, adjusted for the rarity of those words across all documents. Sentences with higher TF-IDF scores are considered to be more significant. *Graph-based Methods (Text Rank),* Sentences are represented as nodes in a graph, with edges linking sentences that include similar words or phrases. The summary includes sentences that score higher on the graph. And *semantic-based technique (LSA),* The singular

value decomposition method is used to find latent semantic links between words and phrases.

In this research, we have experimented with three extractive text summarization techniques namely (summarization using LSA, TF-IDF, and Text ranking). For each technique, we have used an extraction rate of 20%, 30% and 40%. The results of extractive summarization using LSA (Latent Semantic Analysis) leveraged the other.

4.5.1 Latent Semantic Analysis (LSA) and Human Reference Summary (HRS)

An LSA, or Latent Semantic Analysis, is a technique for summarizing and analyzing text. By analyzing the relationships between documents and the terms they contain, it can identify the key concepts and themes in the text based on that information. In this research, we have experimented with three text summarization techniques at the extraction rates of (20%, 30%, and 40%). Figure 12, below indicates a summary evaluation using the LSA technique at a 20% extraction rate using Rouge metrics.

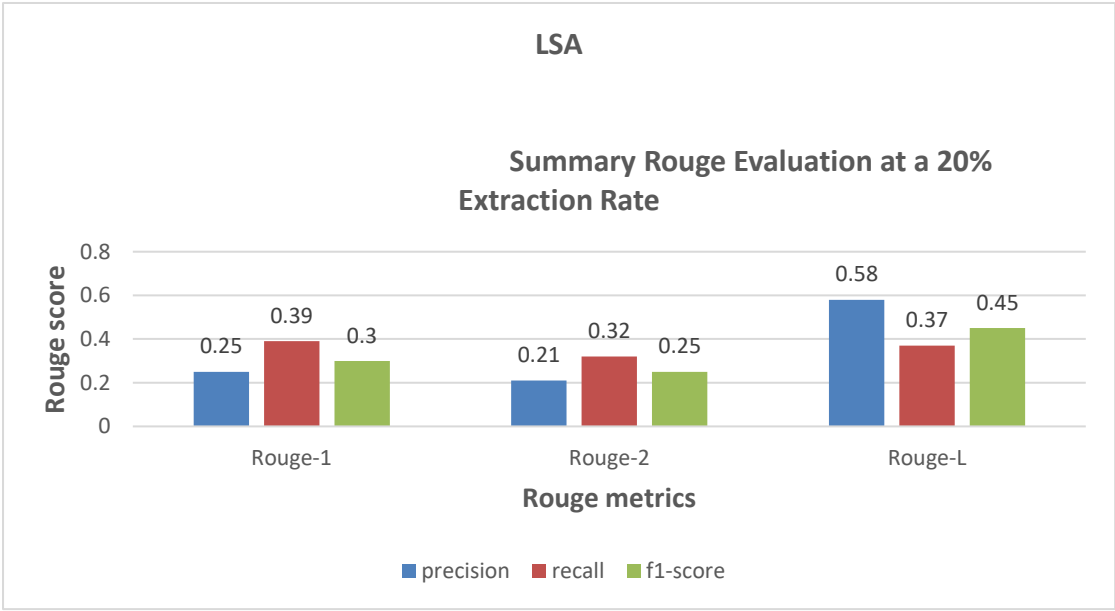


Figure 12: - LSA summary rouge evaluation on a 20% extraction rate

The figure demonstrates rouge evaluation metrics (rouge-1, rouge-2, and rouge-L) in the x-axis and the values (scores) in the y-axis at the extraction rate of 20%. Rouge-1 is all

about the overlapping unigrams, Rouge-2 for bigrams, and Rouge-L for the longest match between human reference summary and machine-generated summary. for each rouge metric the precision, recall, and f1-score are depicted using bars. For each rouge metric, there is a variation in value for example, from rouge-1 to rouge-2 the precision, recall, and f1-score value decreases, whereas from rouge-2 to rouge-L the graph shows an increasing score.

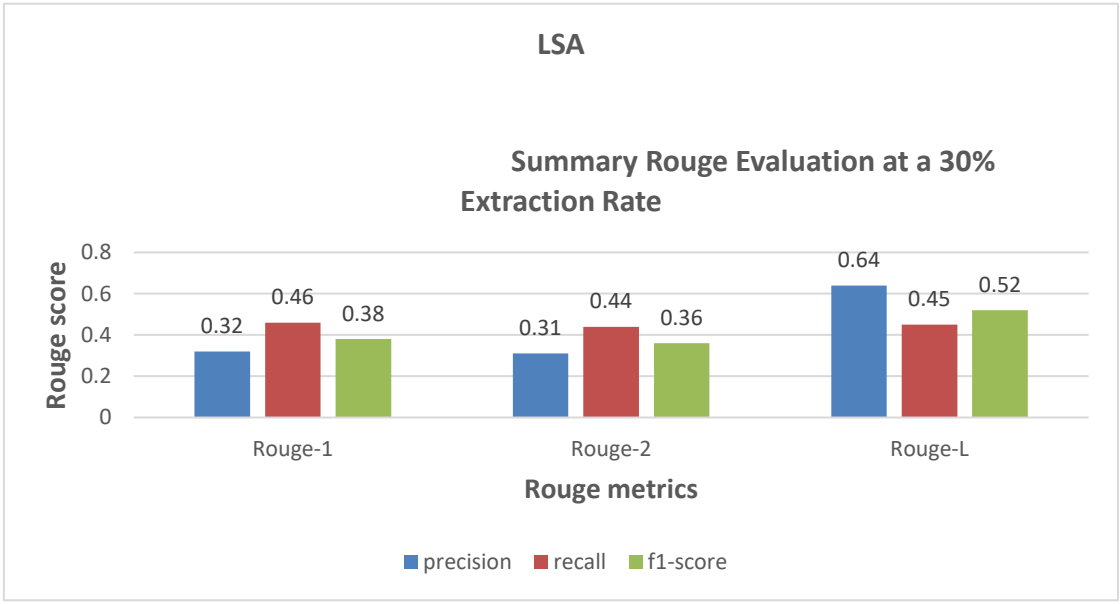


Figure 13: - LSA summary rouge evaluation on a 30% extraction rate

Figure 13 above shows a rouge evaluation metric for an extractive machine-generated summary and an equivalent human reference summary at a 30% extraction rate. As compared to the performance measured in a 20% extraction rate, there is an increase in score for each metric's precision, recall, and f1-score in a 30% extraction rate. Whereas, the same way to a 20% extraction rate a 30% extraction rate also persisted in a score variation.

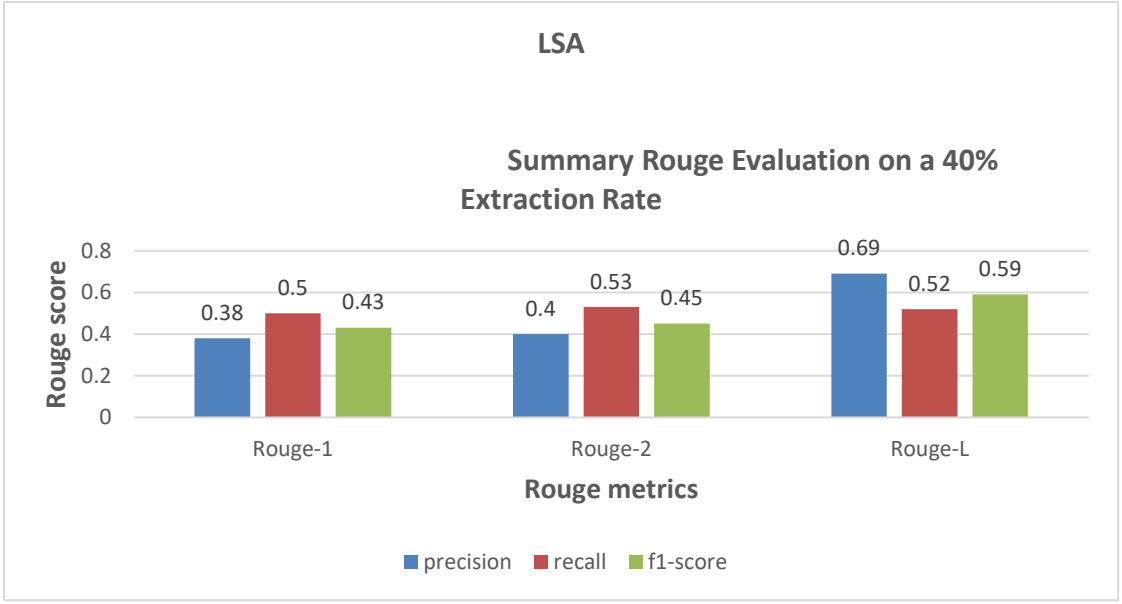


Figure 14: - LSA summary rouge evaluation on a 40% extraction rate

Finally, for the 40% extraction rate, there is an increase in score for each rouge metric's precision, recall, and f1-score as indicated in Figure 14 above.

4.5.2 Text Rank (TR) and Human Reference Summary (HRS)

Text rank is a graph-based ranking model for text processing that is useful for finding keywords and the most relevant sentences in texts. A graph is constructed to find the most relevant sentences within a text, where the vertices represent each sentence, and the edges between sentences are calculated based on how many words they share, namely the number of words they share. So, Figure 15 manifests a rouge metric evaluation for an extractive machine-generated and human reference summary at a 20% extraction rate.

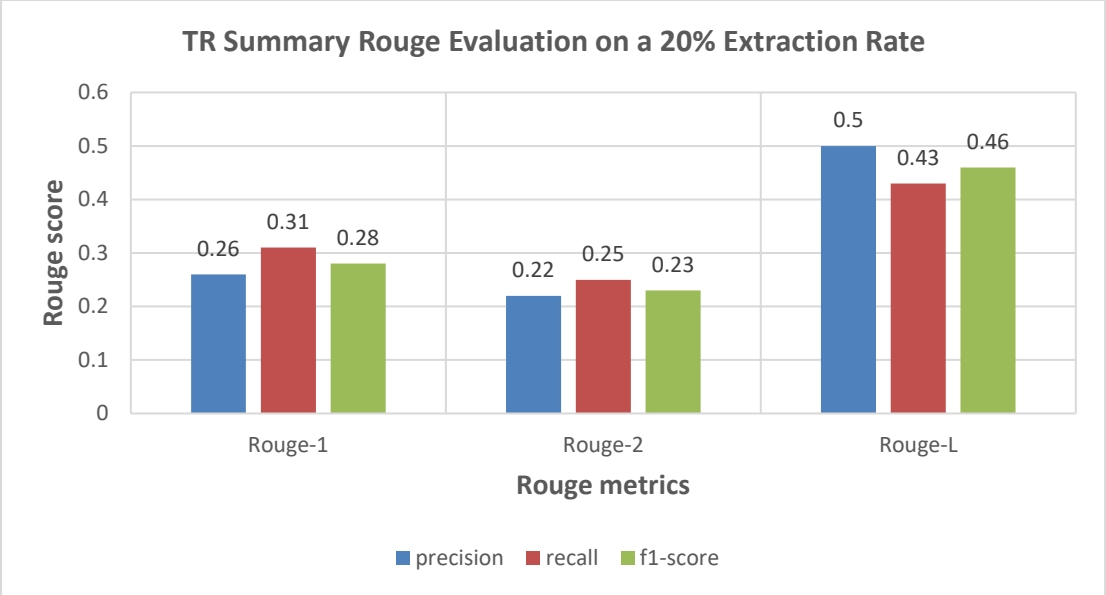


Figure 15: - Text rank summary rouge evaluation on a 20% extraction rate

As shown from the above figure, the precision, recall, and f1-score decrease as we go from rouge-1 to rouge-2 but the inverse is true when we go from rouge-2 to rouge-L the score increases.

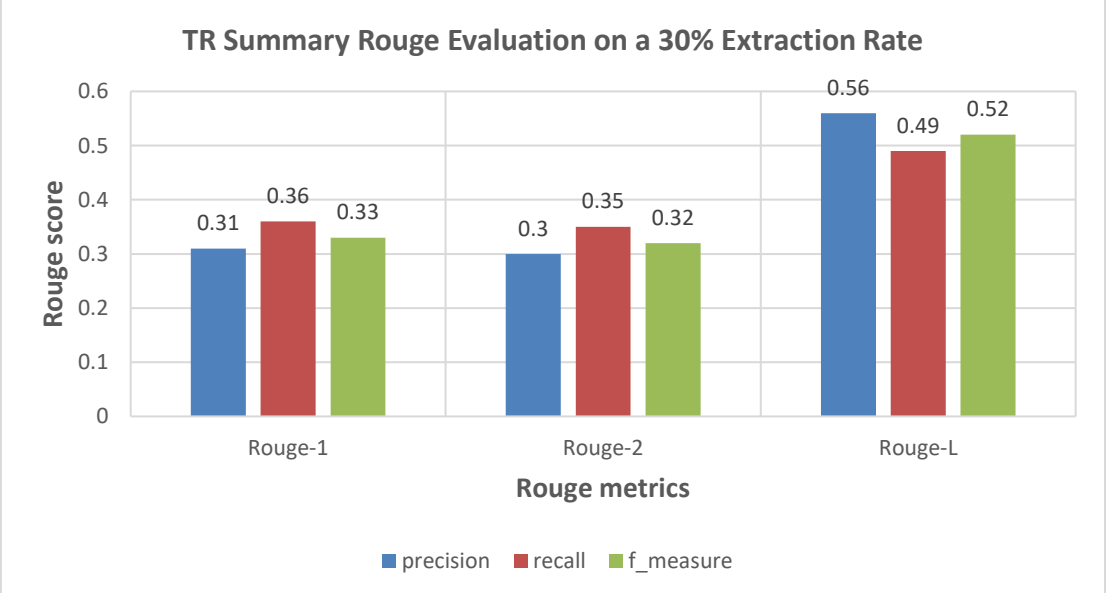


Figure 16: - Text rank summary rouge evaluation on a 30% extraction rate

When we move from rouge-1 to rouge-2 rouge values show a decreasing value in precision, recall, and f1-score.

As compared to text rank summary rouge evaluation on a 20% extraction rate, the 30% extraction rate shows an increased score value in precision, recall, and f1-score for all rouge metrics (rouge-1, rouge-2, and rouge-L).

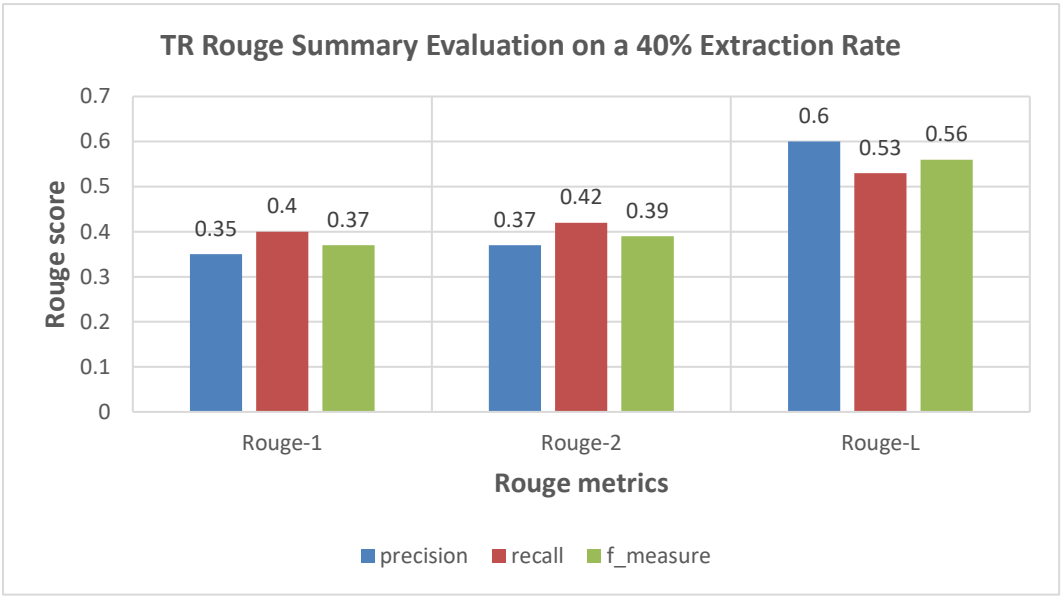


Figure 17: - Text rank summary rouge evaluation on a 40% extraction rate

Text rank summary rouge evaluation on a 40% extraction rate exhibits a steady growth in precision, recall, and f1-score for each rouge evaluation metric (rouge-1, rouge-2, and rouge-L). In the same way as the LSA summary rouge evaluation on a 40% extraction rate, the TR summary rouge evaluation also shows steady growth.

4.5.3 Term Frequency Inverse Document Frequency (TF-IDF) and Human Reference Summary (HRS)

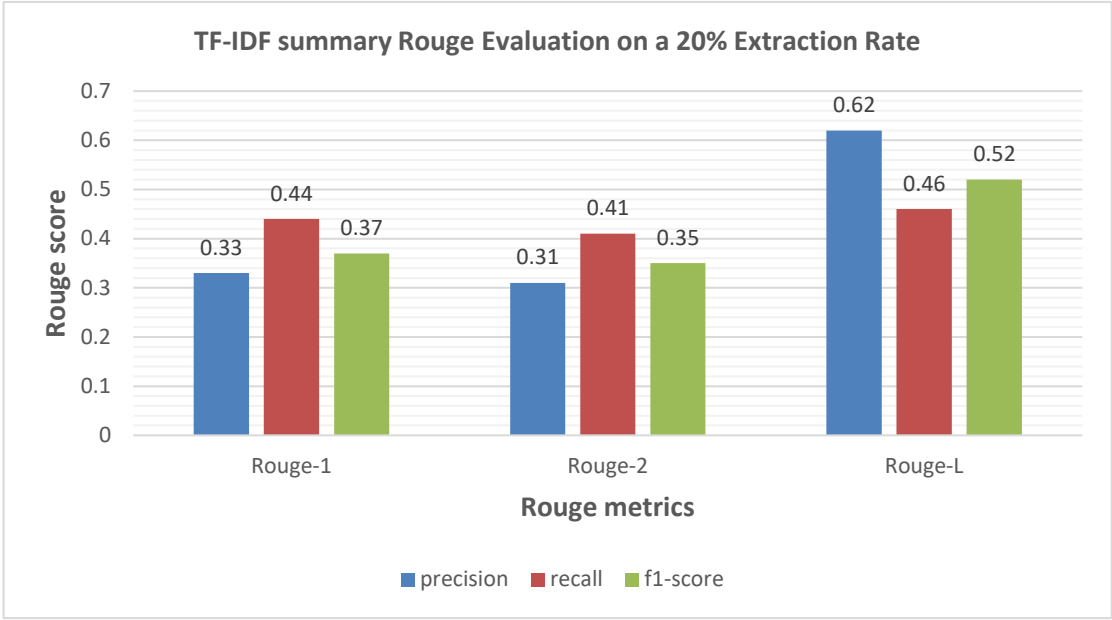


Figure 18: - TF-IDF summary rouge evaluation on a 20% extraction rate

Figure 18 above demonstrates an extractive text summary generated by the machine and a human reference summary with a 20% extraction rate. In the same way as LSA, and TR summary at a 20% extraction rate a TF-IDF summary on a 20% extraction rate shows a decreasing rouge score in precision, recall, and f1-score moving from rouge-1 to rouge-2, while an increasing rouge score moving from rouge-2 to rouge-L.

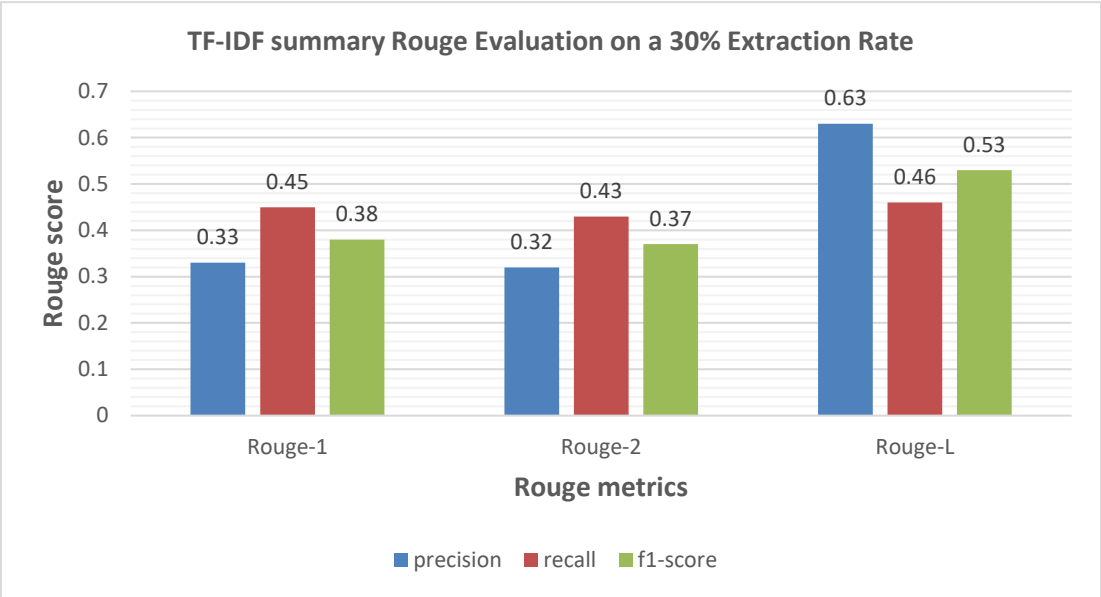


Figure 19: - TF-IDF summary rouge evaluation on a 30% extraction rate

TF-IDF summary on a 30% extraction rate as indicated in Figure 19 showcases, a rouge score dropping from rouge-1 to rouge-2 while an increase moving from rouge-2 to rouge-L.

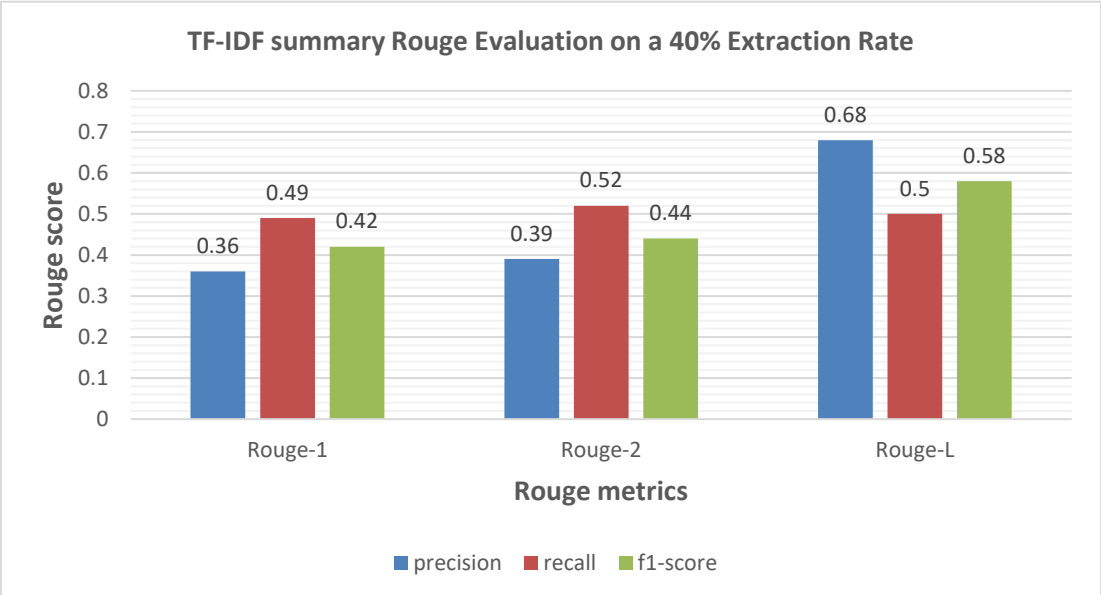


Figure 20: - TF-IDF summary rouge evaluation on a 40% extraction rate

Finally, the TF-IDF summary on a 40% extraction rate shows incremental growth moving from rouge-1 to rouge-2 and then to rouge-L.

4.6 Performance evaluation using cosine similarity

This metric determines the similarity of two non-zero vectors by calculating the cosine of the angle between the two, which ranges from -1 to 1. The closer the cosine similarity is to 1, the more similar the vectors are. As a result, we employ this metric to measure the performance of an extractive summarization. Higher cosine similarity ratings suggest that the summary is more comparable to the source text, which is ideal for a decent summarizing system.

4.6.1 Similarity between Text rank summary and Human reference summary

Table 22: - Similarity between Text rank summary and Human reference summary

TR Vs HRS Similarity in (%)			
	Max	**Min**	**Average**
20% Extraction rate	91	37	66.12
30% Extraction rate	91	51	74.41
40% Extraction rate	90	57	80.17

Table 22 above manifests a cosine similarity between a summary generated by the Text rank algorithm and a human reference summary at the extraction rate of 20%, 30%, and 40%. The max indicates the maximum similarity, the min for the minimum similarity, and the average for the mean. As a result, the maximum average similarity was obtained at the extraction rate of 40%.

4.6.2 Similarity between LSA summary and Human reference summary

Table 23: - Similarity between LSA summary and Human reference summary

LSA Vs HRS Similarity in (%)			
	Max	**Min**	**Average**
20% Extraction rate	89	39	71.15
30% Extraction rate	93	53	79.34
40% Extraction rate	94	66	84.55

The cosine similarity between the summary generated via the LSA technique and the human reference summary indicates a comparably high similarity score than a summary generated by Text rank and TF-IDF techniques. Table 23 depicts the similarity score at the extraction rate of 20%, 30%, and 40%. Similarly, the LSA technique obtains the maximum similarity score at the extraction rate of 40%.

4.6.3 Similarity between TF-IDF summary and Human reference summary

Table 24: - Similarity between TF-IDF summary and Human reference summary

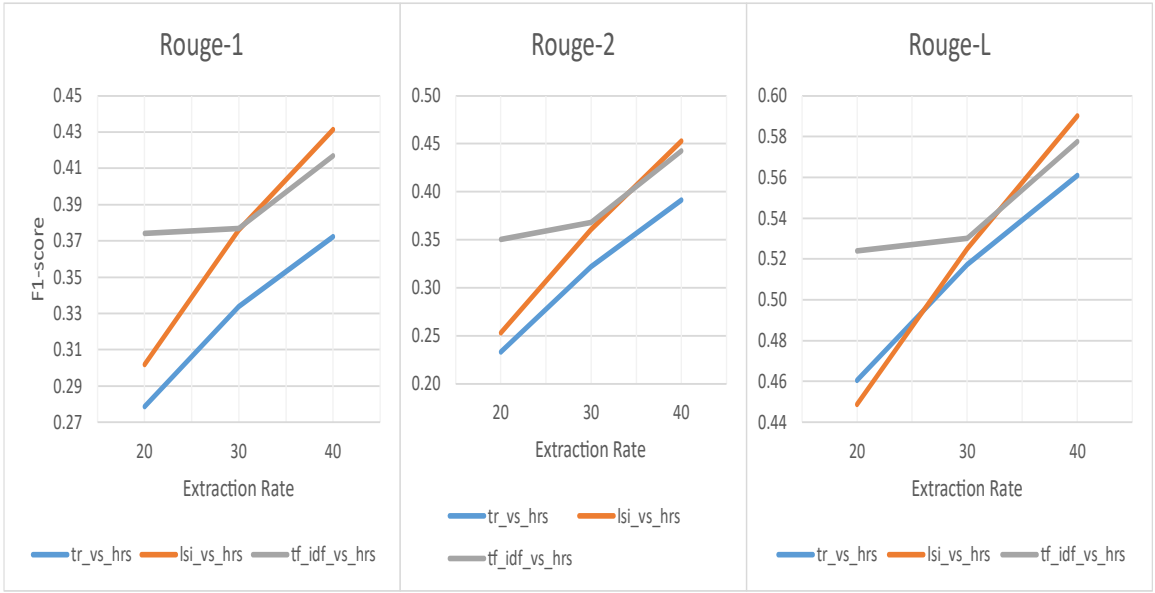
TF-IDF Vs HRS Similarity in (%)			
	Max	**Min**	**Average**
20% Extraction rate	92	37	74.74
30% Extraction rate	92	51	78.66
40% Extraction rate	94	63	84.02

The summary performance evaluation as depicted in Table 24 shows the maximum, minimum, and average similarity on a 20%, 30%, and 40% extraction rate. The same maximum similarity score was obtained for both 20% and 30% extraction rates, but the overall maximum similarity was achieved at 40% extraction rate.

4.7 Summary performance comparison using Rouge for different techniques

An evaluation of a summary's performance is crucial for determining its quality and suitability for real-world use. Even though we have calculated the performance of a machine-generated summary compared to a human-written reference summary, comparing different summarization techniques to determine which one works best remains important. The performance of various summarization methods can be compared to reveal their strengths, weaknesses, and what makes them effective. This comparative analysis can help us make informed decisions about which summarization technique to use for specific use cases or applications.

below demonstrates a summary performance comparison for three techniques (LSA, TR, and TF-IDF). The x-axis shows the extraction rates (20%, 30%, and 40%), while the y-axis shows the f1-score. Furthermore, the figure contains sub-plots. The first sub-plot shows a performance comparison for rouge-1, the second sub-plot shows the performance comparison for rouge-2 and the last for rouge-L.



(a) (b) (c)

Figure 21: - Summary performance comparison using Rouge for different techniques

In Figure 21 (a) above, the tr_vs_hrs (blue line) shows that the F1-score increases steadily from 0.28 at a 20% extraction rate to approximately 0.37 at a 40% extraction rate. In the lsi_vs_hrs (orange line): The F1-score starts at around 0.30 at a 20% extraction rate and increases to about 0.43 at a 40% extraction rate. In the tf_idf_vs_hrs (gray line): F1 scores start at around 0.37 at 20% extraction rates and rise to about 0.42 at 40% extraction rates.

Similarly, in Figure 21 (b), the tr_vs_hrs (blue line) starts at around 0.23 at a 20% extraction rate and increases to about 0.39 at a 40% extraction rate. In the lsi_vs_hrs (orange line): The F1-score starts at around 0.25 at a 20% extraction rate and increases to about 0.45 at a 40% extraction rate. In the tf_idf_vs_hrs (gray line): At a 20% extraction rate, the F1 score starts at around 0.35 and increases to around 0.44 at 40%.

Finally, in Figure 21 (c), tr_vs_hrs (blue line): The F1-score begins at 0.46 at a 20% extraction rate and rises to around 0.56 at a 40% extraction rate. In the lsi_vs_hrs (orange line): The F1-score ranges from 0.44 at a 20% extraction rate to 0.59 at a 40% extraction rate. In the tf_idf_vs_hrs (gray line): At a 20% extraction rate, the F1-score is around 0.52, and at a 40% extraction rate, it is approximately 0.58.

4.8 Summary performance comparison using cosine similarity

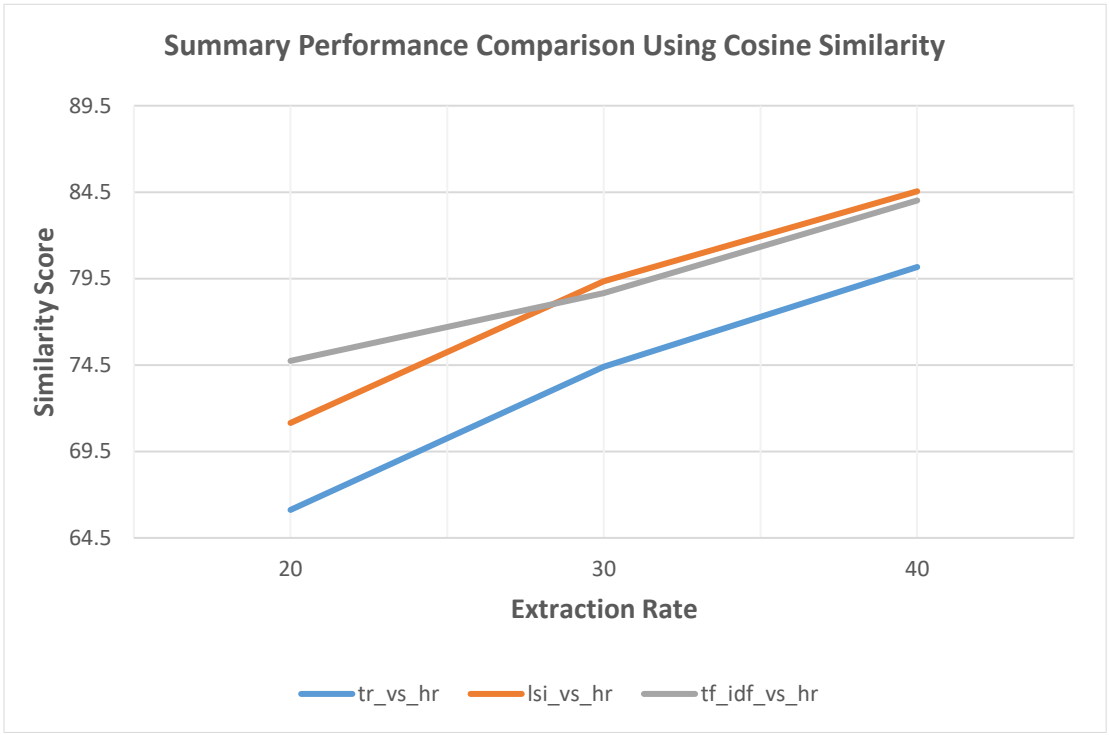


Figure 22: - Summary performance comparison using cosine similarity

Using cosine similarity, Figure 22 compares three different methods (tr_vs_hr, lsi_vs_hr, tf_idf_vs_hr) across three different extraction rates. In the x-axis, the extraction rate is represented by the percentage of text extracted (20%, 30%, 40%). On the other hand, the y-axis represents the similarity score using cosine similarity. In contrast, the lines indicate the cosine similarity at 20%, 30%, and 40% extraction rates.

In the tr_vs_hr graph (blue line): At a 20% extraction rate, the similarity score is around 65.0, and at a 40% extraction rate it's around 75.0. As illustrated by the orange line, the similarity score in lsi_vs_hr increases from 75.0 with a 20% extraction rate to 85.0 with a 40% extraction rate. In tf_idf_vs_hr (gray line), the similarity score starts around 74.5 at a 20% extraction rate and increases to about 84.5 at a 40% extraction rate.

4.9 Discussion of results

The purpose of this research was to design and implement an extractive text summarization system for Awgni news documents. The data sources were the Cherbewa (ቸርቤዋ) newspaper, which was downloaded from the Amhara Media Corporation, as well as hard copies obtained from a shop and library. A total of 213 documents were collected. experts were involved in providing reference summaries equivalent to the source documents. Three different extraction rates were used to generate summaries - 20%, 30%, and 40% of the original text length. The rationale for these specific extraction rates is that a summary should generally not exceed one-third of the source document length, as various factors need to be considered when determining the ideal summary length, such as the purpose of the summarization, the complexity and length of the original text, and the desired level of detail in the summary.

ROUGE (Recall Oriented Understudy for Gisting Evaluation) metrics having precision, recall, and f1-score were used to evaluate the performance of the summary. In addition, a cosine similarity score was also involved.

Our study found that there is a positive correlation between the extraction rate and performance score. Here a correlation indicates a word usage to indicate as the extraction rate increases the performance evaluation score also increases. For each technique (LSA, TR, and TF-IDF) that has been used in our experiment, there is a significant decrease in performance score at the extraction rate of 20% and 30% as we move from rouge-1 to rouge-

L. This is because how the size of a sentence has an effect as we go from rouge-1 (unigrams) and rouge-2 (bi-grams) except for rouge-L.

In this study, a summary comparison was made of different extractive summarization techniques to show which technique outperformed the other. A summary generated via LSA surpasses a summary generated via TR and TF-IDF. Through the use of singular value decomposition (SVD), LSA reduces the dimensionality of a term-document matrix. In this way, the latent semantic structure of a text can be captured, which means words and sentences that share a semantic connection but do not share explicit terms can be identified and grouped.

In summary, the rouge performance evaluation of all the figures in Figure 21 above shows that the summary generated by the LSA summarization technique increasing the extraction rate from 20% to 40% consistently improves the F1 scores. Additionally, in all extraction rates, lsi_vs_hr (orange line) consistently shows the highest similarity scores. The tf_idf_vs_hr (gray line) also performs well, trailing slightly behind lsi_vs_hr, but Tr_vs_hr (blue line) exhibits the lowest similarity scores but increases steadily with the extraction rate.

4.10 An extractive summarization system prototype

Extractive summarization is a type of natural language processing (NLP) technique used to extract sentences, phrases, or key points directly from a given text to create a summary of it. The goal is to produce a concise version of the text containing the most important information and main ideas. To ensure the practical real-world application of an extractive summarization we develop an interactive system. Figure 23, below shows an overall extractive summarization systems user interface.

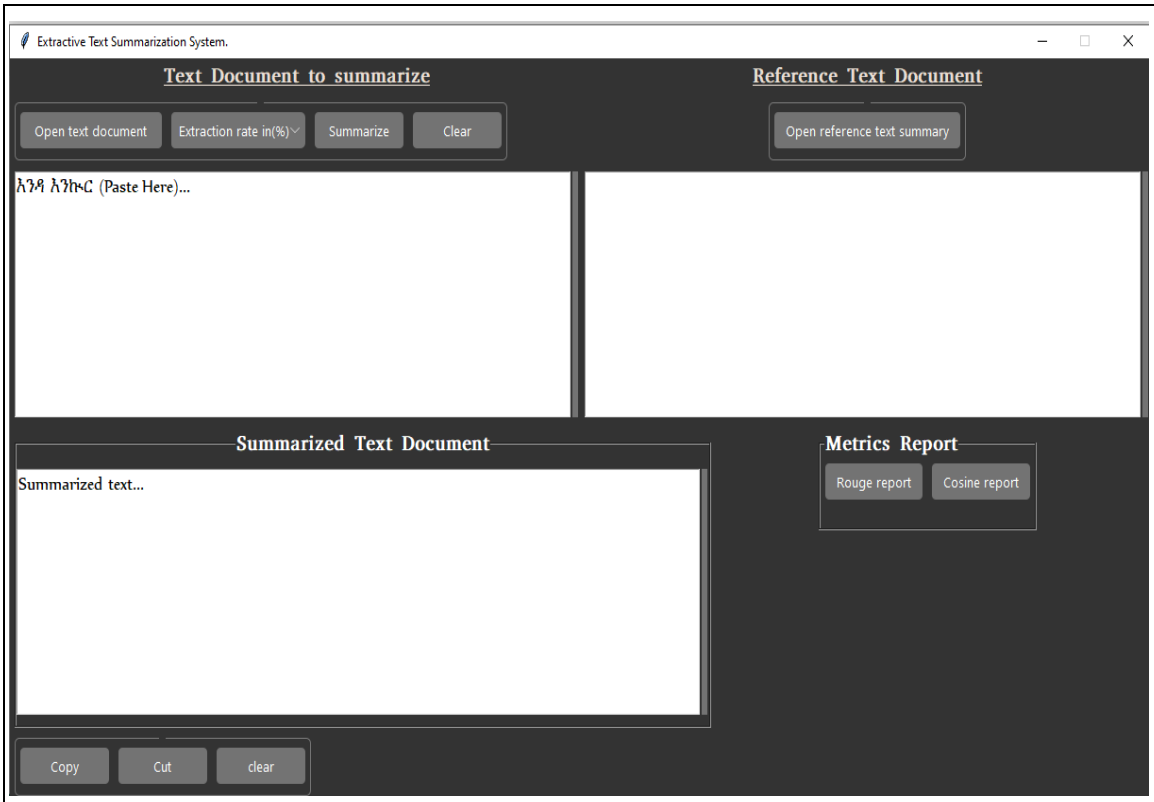


Figure 23: - An extractive summarization system prototype image

The system has different functionalities; at the beginning, the user has to load a text document that has been written in the Awngi language. Then, instantly, the document appears inside the textbox. After that, select the summary size or extraction rate (20%, 30%, or 40%) based on the preference of the user. The following Figure 24 shows how to load an Awngi text document and select the extraction rate (20%) as an example. While doing so, the user can clear the loaded text files to reload another text document.

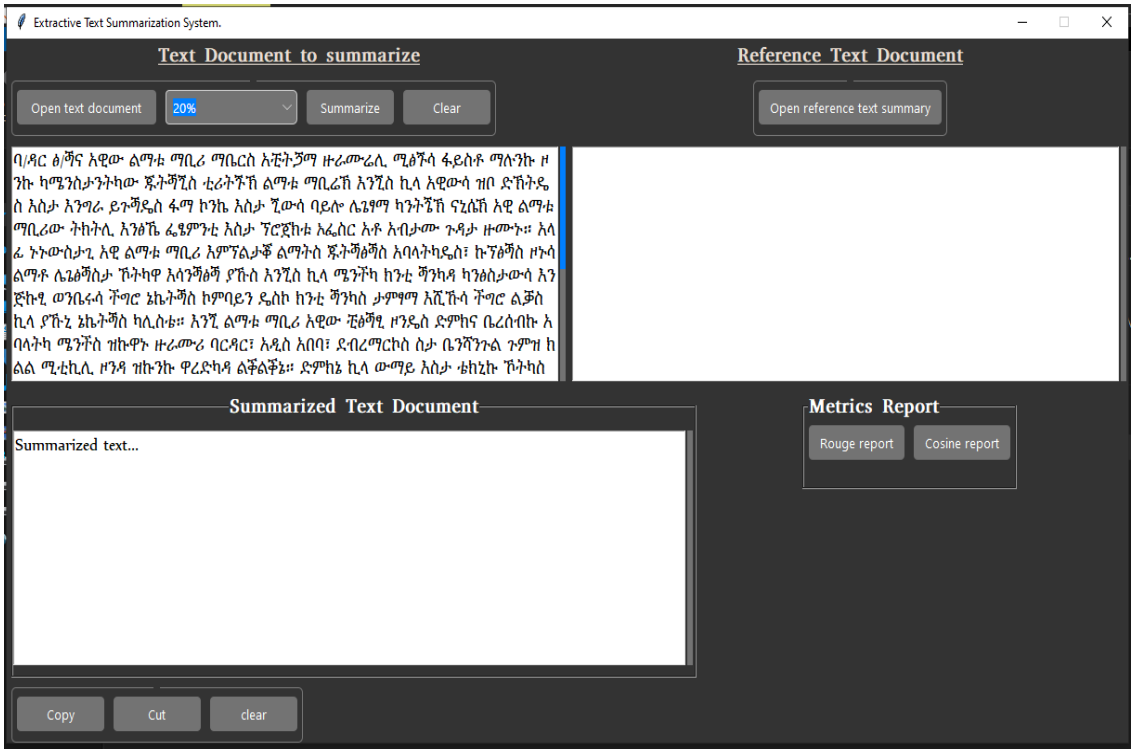


Figure 24: - An image for loading text documents and extraction rate selection

After loading the text document and selecting an appropriate extraction rate the user can click the summarize button, and immediately the summarized text appears inside the summarized textbox at the bottom left corner as indicated in Figure 25 below. In parallel with the summarized text document, there are buttons for copying, cutting, and clearing to make the system interactive and ease of access.

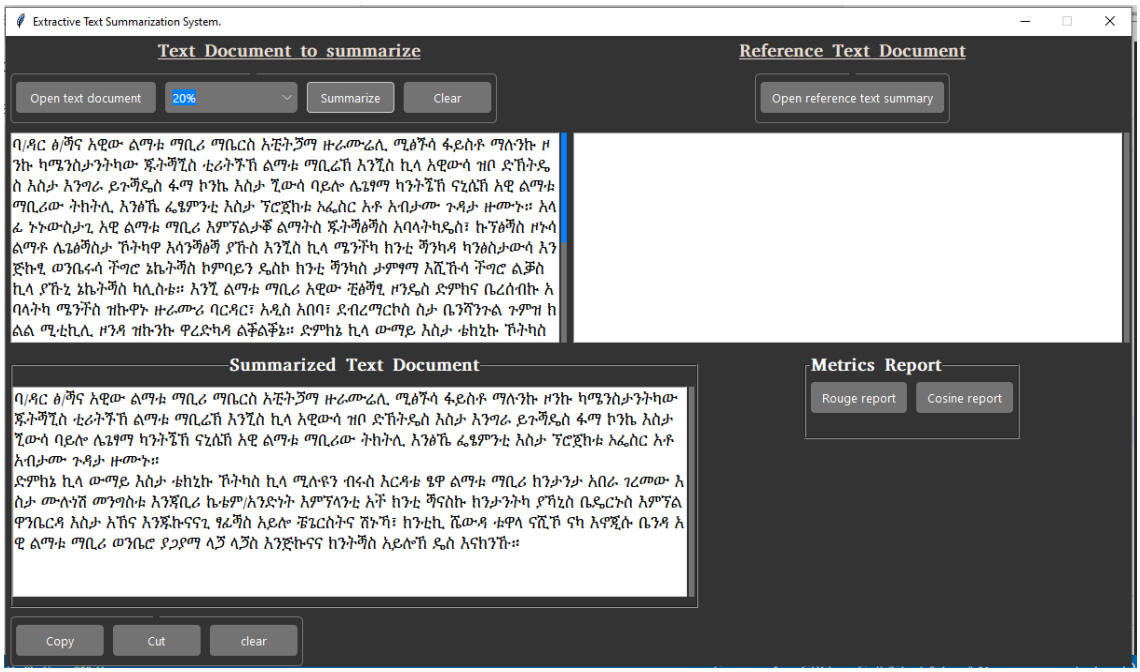


Figure 25: - Sample Summarized text at 20% extraction rate

For the remaining extraction rates (30% and 40%) the user can select iteratively until the last extraction rate which is (40%) based on his or her preference and summarize. Finally, the user can see the performance report of the machine-generated summary with an equivalent human reference summary. The report includes two performance evaluation metrics the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) analysis and cosine similarity. Thereby, the user should have to select an equivalent human reference summary. The following Figure 26 illustrates the whole system beginning from loading a text document to the summary report.

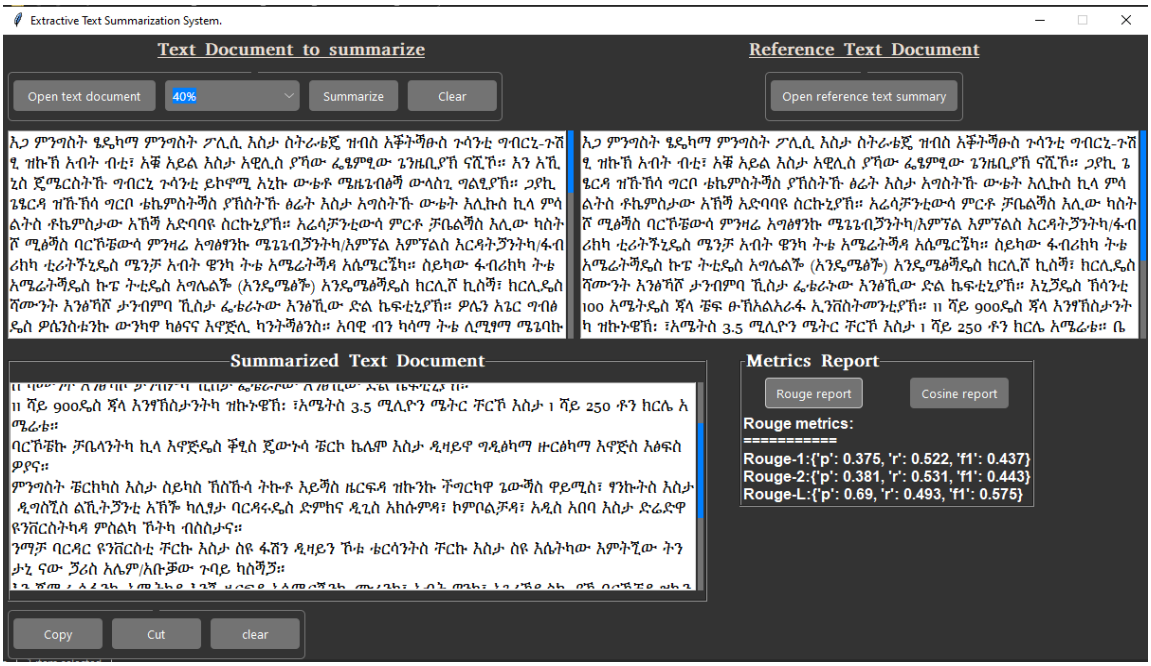


Figure 26: - Performance evaluation using Rouge analysis

Finally, the results show both evaluation metrics Rouge and cosine similarity, each rouge values indicate the matching unigrams in the case of Rouge-1, bigram in the case of Rouge-2, and the longest match of tokens or words in the case of Rouge-L. where the cosine similarity indicates the angle between the machine-generated summary and the human reference summary. The lowest the angle the highest the similarity and the highest the angle the lowest similarity.

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In conclusion, this research work investigated an extractive text summarization system. The process of extracting delineating sentences or paragraphs from an original document and integrating them into a smaller document. For the sake of model development, a total of 213 carefully selected documents were collected from the Amhara Media Corporation and a hard copy of cherbewa (ቸርቤዋ) newspapers from a shop and library.

The purpose of this study was to compare the effectiveness of three extractive text summarizing techniques: LSA (Latent Semantic Analysis), TR (Text Rank), and Term-frequency inverse document frequency (TF-IDF). Our key findings include summarizing Awngi text documents and comparing the results of the summary using performance evaluation techniques such as ROUGE and Cosine similarity metrics. The results indicate that summarization using Latent Semantic Analysis (LSA) outperforms summary generated via Text Rank and TF-IDF, achieving an F1-score of 43.1% for ROUGE-1; an F1-score 45.3% for ROUGE-2; and an F1-score 59.0% for ROUGE-L at the extraction rate of 40% (RQ1).

We assessed the quality of summaries produced at extraction rates of 20%, 30%, and 40% of the original document length through a series of experiments, comparing the summaries to a summary produced by human summarizers. Our analysis of the results shows that the best extraction rate for Awngi news documents is 40%; summaries produced at this extraction level were consistently rated as maintaining the essential information and important details of the source text while being concise enough for readers to easily understand (RQ2).

A collection of 213 thoroughly chosen documents was obtained from Amhara Media Corporation, together with physical copies of cherbewa (ቸርቤዋ) newspapers from a shop and library. Tokenize text documents into sentences, sentences into words then remove stop words, punctuation, an other irrelevant content. Then normalization was made to handle Awngi character encoding. TF-IDF (Term Frequency Inverse Document Frequency) was applied to identify relevant features for extractive summarization of Awngi news documents. A model was developed with a technique having a high performance rate among those compared techniques which was (LSA). A standalone application system was made using python programming language specifically Tkinter module to design a user friendly user interface (RQ3).

We embarked on a thorough evaluation of our text summarization system's performance by contrasting the generated summaries with high-quality human-crafted reference summaries. In order to do this, we first assembled a panel of subject matter experts to produce reference summaries for each source document in our collection of Awngi news documents. We then statistically evaluated how well the produced summaries matched the reference summaries using a variety of automated assessment criteria, such as cosine similarity and ROUGE (RQ4).

Overall, our research shows that an extractive text summarization is feasible for Awngi text documents. So, the effective development of clear and convincing summaries is largely attributed to the combination of a carefully chosen dataset, specific preprocessing methods, and a fine-tuned summarizing model.

Finally, the findings of this work have major significance for many applications in the context of Awngi language content processing, including information retrieval, document summarization, and content aggregation. The study also has significance for autonomous text summarizing applications in the real world across a range of sectors, including

journalism, content creation, and business intelligence. The capacity of the summarization model to handle massive amounts of text and extract crucial insights can greatly improve productivity and decision-making procedures.

5.2 Contributions

The major contributions of this paper include the following: -

- ❖ Provide a literature (fill knowledge gap) for the low-resource Awngi language.
- ❖ Create a practically applicable system (prototype) for an extractive summarization system with a new design and implementation.
- ❖ Create an interactive and easy human reference summary generator system.
- ❖ Compare three different extractive text summarization techniques and give an insight into which technique overperformed the other.
- ❖ In addition to rouge performance metrics, we provide a cosine similarity performance evaluation.

5.3 Recommendations

As a result of this research, an extractive text summarization system has been developed and evaluated for Awngi news documents that have been successfully implemented. Comparing the proposed system to current manual summarization approaches demonstrates the potential for increasing efficiency, reducing time, and reducing costs. The results suggest that automatic text summarization systems have a promising output, but we would like to offer the following recommendations to researchers interested in advancing them: -

Our first recommendation is to incorporate more sophisticated natural language processing techniques, including semantic analysis and discourse modeling, to make generated summaries even more cohesive and informative. Currently, the summary quality

is largely determined by statistical features, and integrating deeper linguistic analysis could enhance it.

In addition, we recommend expanding the dataset used for training and evaluation. Awngi news documents were used in the current study. Expanding the corpus's domain and genre could make the system more generalizable and applicable to a wide range of text.

Furthermore, the approach that was employed in this study was an extractive summarization, so integrating or transforming to an abstractive approach is recommended.

Lastly, we have made a comparative analysis using three base techniques (Text rank, TF-IDF, and Latent Semantic Analysis) including more techniques and testing for overperformance is also another recommendation. Techniques to recommend include, for extractive summarization methods; LexRank:- Similar to TextRank, LexRank uses an eigenvector-based centrality measure to identify important sentences. Additionally, Centroid-based summarization:- This method selects sentences that are most central to the document's topic.

These techniques are recommended based on the following reasons: Expanding the comparative analysis to include more state-of-the-art techniques (e.g., LexRank, Centroid-based), can provide a more comprehensive understanding of the relative strengths and weaknesses of different approaches by utilizing more current techniques. Moreover, the evaluation of overperformance is crucial, for some techniques might perform well on a specific dataset, but do not generalize well to different datasets.

6. REFERENCES

. N. D. (2016). Automatic Text Summarization Using Supervised Machine Learning Technique for Hindi Langauge. *International Journal of Research in Engineering and Technology*, *05*(06), 361–367. https://doi.org/10.15623/ijret.2016.0506065

Agajie, B. A. (2020). Syntactic Object Representations Found in Awgni Sentences. *OKARA: Jurnal Bahasa Dan Sastra*, *14*(1), 99. https://doi.org/10.19105/ojbs.v14i1.3226

Ager, S. (2021). *Rohingya alphabet, pronunciation and language*. Omniglot. https://www.omniglot.com/writing/rohingya.htm

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. In *arXiv preprint arXiv:1707.02268*.

Andhale, N., & Bewoor, L. A. (2017). An overview of text summarization techniques. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*. https://doi.org/10.1109/ICCUBEA.2016.7860024

Azhari, M., & Jaya Kumar, Y. (2017). Improving text summarization using neuro-fuzzy approach. *Journal of Information and Telecommunication*, *1*(4), 367–379. https://doi.org/10.1080/24751839.2017.1364040

Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K., & Ghosh, S. (2021). Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3462757.3466092

Computing, F. O. F., & Guadie, B. Y. A. (2017). *Amharic Text Summarization for News Items posted on Social Media This Thesis Submitted to Faculty of Computing to Jimma University for the Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Technology Amharic Text Su*.

Demis, T. D. (2018). *Automatic Amharic Multi document News Text Summarization using Open Text Summarizer*.

Dinegde, G. D., & Tachbelie, M. Y. (2014). Afan Oromo News Text Summarizer. *International Journal of Computer Applications*, *103*(4), 975–8887.

Dingare, A., Bein, D., Bein, W., & Verma, A. (2022). *Abstractive Text Summarization Using Machine Learning*. 269–276. https://doi.org/10.1007/978-3-030-97652-1_33

Document Summarization Using Latent Semantic Indexing | by Srinivas Chakravarthy | Towards Data Science. (n.d.). Retrieved April 23, 2024, from https://towardsdatascience.com/document-summarization-using-latent-semantic-indexing-b747ef2d2af6

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/10.1016/j.eswa.2020.113679

Fiori, A. (2019). Trends and applications of text summarization techniques. In *Trends and Applications of Text Summarization Techniques*. https://doi.org/10.4018/978-1-5225-9373-7

Garbade, Dr. M. J. (2018). *A Quick Introduction to Text Summarization in Machine Learning | by Dr. Michael J. Garbade | Towards Data Science*. https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

Guadie, A., Tesfaye, D., & Kebebew, T. (2021). Amharic Text Summarization for News Items Posted on Social Media. *International Journal of Intelligent Information Systems*, *10*(6), 125. https://doi.org/10.11648/j.ijiis.20211006.14

Gupta, M., & Garg, N. K. (2016a). Text summarization of Hindi documents using rule based approach. *Proceedings - 2016 International Conference on Micro-Electronics and Telecommunication Engineering, ICMETE 2016*, 366–370. https://doi.org/10.1109/ICMETE.2016.104

Gupta, M., & Garg, N. K. (2016b). Text summarization of Hindi documents using rule based approach. *Proceedings - 2016 International Conference on Micro-Electronics and Telecommunication Engineering, ICMETE 2016*, 366–370. https://doi.org/10.1109/ICMETE.2016.104

Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, *2*(3), 258–268. https://doi.org/10.4304/jetwi.2.3.258-268

Hetzron, R. (1978). The nominal system of awngi (southern agawxs). *Bulletin of the School of Oriental and African Studies*, *41*(1), 121–141. https://doi.org/10.1017/S0041977X00057815

Issam, K. A. R., Patel*, S., & N., S. C. (2020). Topic Modeling Based Extractive Text Summarization. *International Journal of Innovative Technology and Exploring Engineering*, *9*(6), 1710–1719. https://doi.org/10.35940/ijitee.f4611.049620

Joshi, M. L., Joshi, N., & Mittal, N. (2021). SGATS: Semantic Graph-based Automatic Text Summarization from Hindi Text Documents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(6), 1–32. https://doi.org/10.1145/3464381

Joy Winnie Wise, D. C., Ambareesh, S., Ramesh, B. P., Sugumar, D., Bhimavarapu, J. P., & Senthil Kumar, A. (2024). Latent Semantic Analysis Based Sentimental Analysis of Tweets in Social Media for the Classification of Cyberbullying Text. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(7s), 26–35.

Juan-Manuel, T.-M., Manuel, J., & Moreno, T. (2014). *Summarization*.

Kerui, Z., Haichao, H., & Yuxia, L. (2020). Automatic text summarization on social media. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3440084.3441182

Krishnakumar, A., Naushin, F. A. R., Mrithula, K. L., & Bharathi, B. (2022). Text summarization for Indian languages using pre-trained models. *CEUR Workshop Proceedings*, *3395*, 424–434.

Languages of Ethiopia. (2021). https://liupalmer.libguides.com/c.php?g=1143777&p=8348034

Latent Semantic Analysis (LSA). (2023). https://blog.marketmuse.com/glossary/latent-semantic-analysis-definition/#:~:text=Latent Semantic Analysis is a,relationships between terms and concepts.

Li, A., Jiang, T., Wang, Q., & Yu, H. (2016). The mixture of TextRank and LexRank techniques of single document automatic summarization research in Tibetan. *Proceedings - 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2016*, *1*, 514–519. https://doi.org/10.1109/IHMSC.2016.278

Mattupalli*, S., Bhandari, A., & Praveena, B. J. (2020). Text Summarization using Deep Learning. *International Journal of Recent Technology and Engineering (IJRTE)*, *9*(1), 2663–2667. https://doi.org/10.35940/ijrte.a3056.059120

Patil, S. R. (2011). *Domain Specific e-Document Summarization Using Extractive Approach*.

Patil, V., Krishnamoorthy, M., Oke, P., & Kiruthika, P. M. (2020). *A Statistical Approach for Document Summarization*. 33–43.

Pattnaik, S., & Nayak, A. K. (2019). Summarization of odia text document using cosine similarity and clustering. *Proceedings - 2019 International Conference on Applied Machine Learning, ICAML 2019*, 143–146. https://doi.org/10.1109/ICAML48257.2019.00035

Protim Ghosh, P., Shahariar, R., & Hossain Khan, M. A. (2018). A Rule Based Extractive Text Summarization Technique for Bangla News Documents. *International Journal of Modern Education and Computer Science*, *10*(12), 44–53. https://doi.org/10.5815/ijmecs.2018.12.06

Redi, A. (2020). *Extractive Amharic Text Summarization Using Latent Dirichlet Allocation*. *September*.

Reeve, L. H., Han, H., & Brooks, A. D. (2007). The use of domain-specific concepts in biomedical text summarization. In *Information Processing and Management* (Vol. 43, Issue 6, pp. 1765–1776). https://doi.org/10.1016/j.ipm.2007.01.026

Saggion, H., & Poibeau, T. (2013). *Automatic Text Summarization: Past, Present and Future*. 3–21. https://doi.org/10.1007/978-3-642-28569-1_1

Shiferaw, S. (2015). *Awi Tourist guide*.

Sri, S. H. B., & Dutta, S. R. (2021). A survey on automatic text summarization techniques. *Journal of Physics: Conference Series*, *2040*(1), 121–135. https://doi.org/10.1088/1742-6596/2040/1/012044

Sri, T., Raju, R., & Allarpu, B. (2017). Text Summarization using Sentence Scoring Method. *International Research Journal of Engineering and Technology*, 2395–56.

Tashoma, L., Advisor, F., & Assabie, Y. (2020a). *Afaan Oromo Text Summarization using Word Embedding Lamesa Tashoma Fanache*.

Tashoma, L., Advisor, F., & Assabie, Y. (2020b). *Afaan Oromo Text Summarization using Word Embedding Lamesa Tashoma Fanache*.

Taunk, D., & Varma, V. (2022). Summarizing Indian Languages using Multilingual Transformers based Models. *CEUR Workshop Proceedings*, *3395*, 435–442.

Torres Moreno, J. M. (2014). Automatic Text Summarization. *Automatic Text Summarization*, *9781848216*, 1–348. https://doi.org/10.1002/9781119004752

Wazery, Y. M., Saleh, M. E., Alharbi, A., & Ali, A. A. (2022). Abstractive Arabic Text Summarization Based on Deep Learning. *Computational Intelligence and Neuroscience*, *2022*, 1–14. https://doi.org/10.1155/2022/1566890

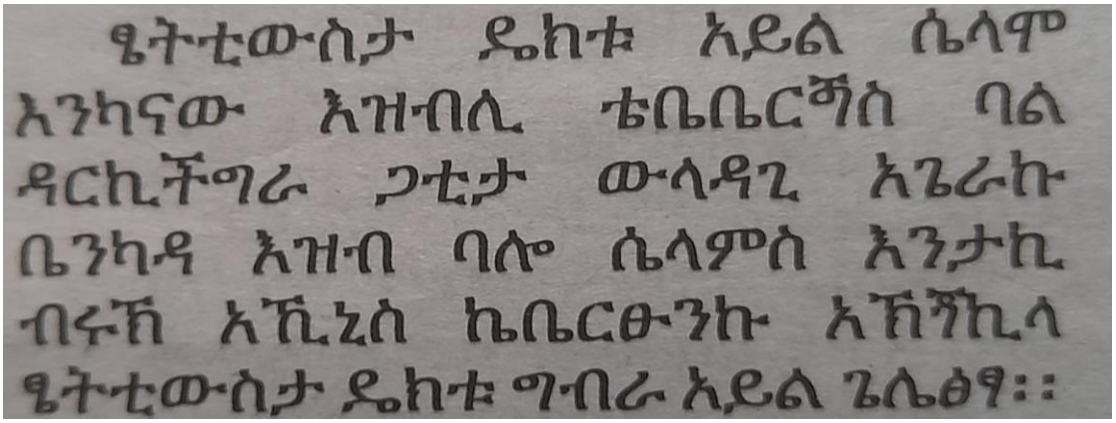
Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Yirdaw, E. D., & Ejigu, D. (2012). Topic-based amharic text summarization with probabilisic latent semantic analysis. *Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES 2012*, 8–15. https://doi.org/10.1145/2457276.2457279

Zaki, A. M., Khalil, M. I., Abbas, H. M., Supreetha, D., Rajeshwari, S. B., & Kallimani, J. S. (2020). Abstractive Text Summarization. *Journal of Xidian University*, *14*(6), 26884–26888. https://doi.org/10.37896/jxu14.6/094

APPENDICES

APPENDICES A: - SAMPLE DOCUMENT CONVERSION FROM IMAGE TO TEXT USING GOOGLE SHEET



ቄትቲውስታ ዴክቱ አይል ሴላም እንካናው እዝብሊ ቴቤቤርሻስ ባል ዳርኪችግራ ጋቲታ ውላዳጊ አጌራኩ ቤንካዳ እዝብ ባሎ ሴላምስ እንታኪ ብሩክ አኺኒስ ኬቤርፀንኩ አክኝኪላ ቄትቲውስታ ዴክቱ ግብራ አይል ጌሌፅፃ።

APPENDICES B: - STOP WORDS LIST

Table 25: - List of stop words

አን	ይው	እኖጂ	እኖጂሱ
ኚ	እንት	እንቶጂሱ	ኩው
ኚዉ	ጋጂሱ	እስታ	ያኼስጉ
አኹኪ	ምክናያትኪ	እኒ	እን
አኒ	እንዳ	ኚዳ	አንዳ
እስታ	ኪላ	አክኼ	አክሻስ
ዬሜካ	ኺስታ		

APPENDICES C: - MANUAL SUMMARY PREPARATION GUIDELINE

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

Program: -MSc in Computer Science

TITLE: - DESIGN AND IMPLEMENTATION OF AN EXTRACTIVE TEXT SUMMARIZATION SYSTEM FOR AWNGI (አውኒ) NEWS DOCUMENTS

Summary request:

The aim of this study is to design and implement an extractive text summarization system for Awngi (አውኒ) news documents. In the age of the digital world, we are constantly bombarded by vast amounts of information overload. As a result, summarization plays a significant role in navigating through this information overload by condensing lengthy texts, articles, or documents into concise summaries. Through it, we can quickly grasp the main points and key information without having to read or process the entire text in its entirety. Although automatic (machine-generated) summaries are invaluable in condensing large amounts of text in a fraction of the time, they need equivalent human-annotated reference summaries to evaluate the performance of machine-generated summaries. So, we kindly request your expertise in ranking the given Awngi news documents.

To do so, we have prepared an interactive system to facilitate the work and reduce the total amount of time taken. The system has a user interface with four functionalities. The first functionality allows the expert to select a source document in the form of a drop-down menu. While selecting the source document the sentences in the document are displayed with a checkbox to select (the second functionality).

Before selecting any of the sentences please select an extraction rate (20%) which is the third functionality. Then select a checkbox having a sentence considering informativeness, non-redundancy, coverage, and coherence. Finally, click the save button after reaching the maximum extraction rate (it will automatically disable the checkboxes). This process will proceed until a 40% extraction rate.

1. Informativeness:

A summary's informativeness is determined by the extent to which it captures all of the important ideas, main points, and details from the source text.

2. Non-redundant:

A non-redundant summary avoids duplication of information or unnecessary repetition of material. It presents the essential content of the original text without repeating it.

3. Coverage:

Coverage describes how well a summary includes all the critical information described in the original text. It should cover all the key points, key details, and essential aspects of its source material.

4. Coherence:

A coherent summary provides readers with a clear understanding of the narrative or argument through logical flow, organization, and smooth progression of ideas.

Kind regards,

Muluken Tilahun

May 21, 2015 E.C

APPENDICES D: - DOCUMENT STATISTICS

	document	clean_doc	#sen	#words_per_sen	#words_per_doc	max_words_per_sen	min_words_per_sen	avg
0	[እጋ ምንግስት ፔዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቕትኻውስ ...	[እጋ ምንግስት ፔዴካማ ምንግስት ፖሊሲ ስትራቴጂ ዝብስ አቕትኻ ጉሳን ግብ...	30	[22, 11, 17, 19, 20, 8, 14, 8, 18, 13, 18, 9, ...	466	31	5	15.533333
1	[ታፎ እስታ ልኮ አፄፍኻስ እንጅኩካማ ኾ ታብሉሳ፣ ዛኻ ዝሜድካውስ ታፎ ካ...	[ታፎ ልኮ አፄፍኻስ እንጅኩካማ ታብሉሳ ዛኻ ዝሜድካውስ ታፎ ካንትኻዲስ ፌ...	45	[21, 23, 13, 12, 11, 14, 13, 42, 15, 8, 6, 13,...	699	42	5	15.533333
2	[ባይል ይወትስ ዝኩኑስ እኑስ እኑዋ ካሲትኻናው፣ እምኘልዳ ኹናው ዝቕናውሳ...	[ባይል ይወትስ ዝኩኑስ እኑስ እኑዋ ካሲትኻናው እምኘልዳ ኹናው ዝቕናውሳ ...	43	[30, 24, 22, 11, 8, 12, 13, 19, 17, 12, 23, 36...	894	42	5	20.790698
3	[ውላዴስ ቤዴርትስ ደራሲ (ፃፋንቲ) ሊሊቲ ሼው ንኹ ካላአቕት ዝኩኽ አኽኹ...	[ውላዴስ ቤዴርትስ ደራሲ ፃፋንቲ ሊሊቲ ሼው ንኹ ካላአቕት ዝኩኽ ሼውንኻ ...	32	[12, 27, 12, 8, 21, 7, 8, 8, 10, 5, 11, 8, 9, ...	353	30	5	11.031250
4	[ከርፂ ሺፅኻፂውሳ ችግሮ ኽይኻስ ጂሜርስትኹ ልቕልቕቲ ሲፋታ ኩርፄንት እን...	[ከርፂ ሺፅኻፂውሳ ችግሮ ኽይኻስ ጂሜር ልቕልቕቲ ሲፋታ ኩርፄንት እንፃኽስ...	40	[11, 30, 29, 13, 15, 20, 15, 15, 29, 14, 31, 2...	811	35	8	20.275000

APPENDICES E: - HUMAN REFERENCE SUMMARY AT 20%, 30%, and 40%

EXTRACTION RATE

	Rf_20%_sum	Rf_30%_sum	Rf_40%_sum
0	[ኢጋ ምንግስት ፔዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻውስ ...	[ኢጋ ምንግስት ፔዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻውስ ...	[ኢጋ ምንግስት ፔዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻውስ ...
1	[ወታትካ ጋጻ ዝኩኸሳ ግርቦ እስታ ጌንዜቦ ቱኬምስትሻስ ጋውሳ አከሞ አጌራ...	[ወታትካ ጋጻ ዝኩኸሳ ግርቦ እስታ ጌንዜቦ ቱኬምስትሻስ ጋውሳ አከሞ አጌራ...	[ወታትካ ጋጻ ዝኩኸሳ ግርቦ እስታ ጌንዜቦ ቱኬምስትሻስ ጋውሳ አከሞ አጌራ...
2	[እዊ እዝብ ሺጋርትጋሱ ዴሶ፣ ባይል እስታ ቺፅሻሂው እኺኒ ዝኩኸ ያኹስ እ...	[ባይል ይወትስ ዝኩኑስ እኑስ እኑዋ ካሲትሻናው፣ እምኘልጻ ኹናው ዝቾናውሳ...	[ባይል ይወትስ ዝኩኑስ እኑስ እኑዋ ካሲትሻናው፣ እምኘልጻ ኹናው ዝቾናውሳ...
3	[ውላዴስ ቤዴርትስ ደራሲ (ፃፋንቲ) ሊሊቲ ሼው ንኾ ካላአቾት ዝኩኸ አኸኼ...	[ውላዴስ ቤዴርትስ ደራሲ (ፃፋንቲ) ሊሊቲ ሼው ንኾ ካላአቾት ዝኩኸ አኸኼ...	[ውላዴስ ቤዴርትስ ደራሲ (ፃፋንቲ) ሊሊቲ ሼው ንኾ ካላአቾት ዝኩኸ አኸኼ...
4	[ንማቻ ባርጻር ኬቴምጻ አማኻሪ ክልሉ ኩስኺኻናጻ እዝብኩ ዋኬልስታንትካው ...	[ከርዊ ቺፅሻዊውሳ ችግሮ ኸይቻስ ጀሜርስትኹ ልቾልቾቲ ሲፋታ ኩርዜንት እን...	[ከርዊ ቺፅሻዊውሳ ችግሮ ኸይቻስ ጀሜርስትኹ ልቾልቾቲ ሲፋታ ኩርዜንት እን...
5	[ከርዊ ቺፅሻዊውሳ ችግሮ ፌቴፋንኩ አመራርካ እስታ ውማይቴንካ ካሲስታንትካ...	[ከርዊ ቺፅሻዊውሳ ችግሮ ፌቴፋንኩ አመራርካ እስታ ውማይቴንካ ካሲስታንትካ...	[ከርዊ ቺፅሻዊውሳ ችግሮ ፌቴፋንኩ አመራርካ እስታ ውማይቴንካ ካሲስታንትካ...
6	[ያኼስጉ እኾሺው ምርቱ ዜርፍስ ዜሬቾቲኔ ምርቶ አሜሬታንኩ አጌርካ ክቻዴስ...	[ያኼስጉ እኾሺው ምርቱ ዜርፍስ ዜሬቾቲኔ ምርቶ አሜሬታንኩ አጌርካ ክቻዴስ...	[እንሴስካው አብቱ ልማትስ ኬሸኹሳ ቼፎ እሚሻስ አሌምዴስ 10ንቱሳ አፍሪካ...
7	[14ንቲ ሊኸ ዝቤን ቤናዴስ ጀሜርሻስ ንጉስካ ባርጻሮ እንካናሺካ፡፡ ስሜን...	[ባርጻር 1922 ም/እ ትዴስ ጀሜርሻስ አቾቶ አግዔታጊ ቲንቱት ኬቴማሸ፡፡,...	[ባርጻር 1922 ም/እ ትዴስ ጀሜርሻስ አቾቶ አግዔታጊ ቲንቱት ኬቴማሸ፡፡,...
8	[እዝብ ሴላምስ ፋማ ሴላምስ ቱዋታ፣ ጊፅሳማ ፊቲሳታ፣ ከንታማ ጉድ አችጻ ...	[እዝብ ሴላምስ ፋማ ሴላምስ ቱዋታ፣ ጊፅሳማ ፊቲሳታ፣ ከንታማ ጉድ አችጻ ...	[እዝብ ሴላምስ ፋማ ሴላምስ ቱዋታ፣ ጊፅሳማ ፊቲሳታ፣ ከንታማ ጉድ አችጻ ...

APPENDICES F: - MACHINE-GENERATED SUMMARY AT DIFFERENT EXTRACTION RATE FOR DIFFERENT TECHNIQUES

TR_20%_sum	TR_30%_sum	TR_40%_sum	lsi_20%_sum	lsi_30%_sum	lsi_40%_sum	tf_idf_20%_sum	tf_idf_30%_sum	tf_idf_40%_sum
[አሬሳቻንቲውሳ ምርቶ ቻቤልሻስ እሊው ካስትሾ ሚፅሻስ ባርኾቼውሳ ምንዛሬ ...	[ጋያኪ ጌዬርዳ ዝኹኽሳ ግርቦ ቴኬምስትሻስ ያኽስትኹ ፅሬት እስታ አግስትኹ...	[ጋያኪ ጌዬርዳ ዝኹኽሳ ግርቦ ቴኬምስትሻስ ያኽስትኹ ፅሬት እስታ አግስትኹ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...	[እጋ ምንግስት ዬዴካማ ምንግስት ፖሊሲ እስታ ስትራቴጂ ዝብስ አቾትሻፁስ ...
[1999 ም/አ ታዳ ታውሊ እንፅኻሾ ቷኽ አኽች ታከስፅኹ ወታት ፀሃይ እን...	[ወታት ፀሃይ አድማሱሊ ታምትሻናና ሺው አችትዳ አኹኪ እንፃኽፂ ስፍሪዳ እ...	[ወታት ፀሃይ አድማሱሊ ታምትሻናና ሺው አችትዳ አኹኪ እንፃኽፂ ስፍሪዳ እ...	[ወታትካ ሻዳ ዝኩኽሳ ግርቦ እስታ ጌንዜቦ ቴኬምስትሻስ ሻውሳ አከሞ አጌራ...	[ወታትካ ሻዳ ዝኩኽሳ ግርቦ እስታ ጌንዜቦ ቴኬምስትሻስ ሻውሳ አከሞ አጌራ...	[ወታትካ ሻዳ ዝኩኽሳ ግርቦ እስታ ጌንዜቦ ቴኬምስትሻስ ሻውሳ አከሞ አጌራ...	[ታፎ እስታ ልኮ አፄፍሻስ እንጅኩካማ ቾ ታብሉሳ፣ ዛኻ ዝሜድካውስ ታፎ ካ...	[ታፎ እስታ ልኮ አፄፍሻስ እንጅኩካማ ቾ ታብሉሳ፣ ዛኻ ዝሜድካውስ ታፎ ካ...	[ታፎ እስታ ልኮ አፄፍሻስ እንጅኩካማ ቾ ታብሉሳ፣ ዛኻ ዝሜድካውስ ታፎ ካ...