Afan Oromo Text Summarization With Deep Learning Approach

A thesis submitted in partial fulfillment for the degree of
Masters in Computer Science

To

The Faculty of Informatics

Of

St. Mary's University

By
Hawi Tamiru Geleta

Addis Ababa
Ethiopia

February 15, 2022

ACCEPTANCE

Afan Oromo Text Summarization With Deep Learning Approach

By

Hawi Tamiru Geleta

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirement for the degree of Masters of Science in Computer Science

Thesis Examination Committee:

____

Internal Examiner

Sileshi Demesie (PHD)

External Examiner

____

Dean, Faculty of Informatics

February 15, 2022

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Hawi Tamiru Geleta

____
Signature

Addis Ababa
Ethiopia

This thesis has been submitted for examination with my approval as advisor.

____
Signature

Addis Ababa

Ethiopia

February 15, 2022

ACKNOWLEDGEMENT

First and foremost I would thank the Almighty God who has led me in success through this project in particular and throughout my life in general. My gratitude also goes to my advisor Mr. Hafte Abera for his invaluable sincere advice and guidance to complete the work. Besides, I appreciate my father, mother and brother whose jointly family support has strengthened me from the start to the end. It is also my pleasure to extend my appreciation to my ever trusted friend Helen Akelew who had reinforced my courage with her wise advice. Last, but not least my due thanks goes to my freind Kasahun Abera for his consistent support and encouragement.

Table of Contents

ACRONYMS

AM	Attention mechanisms
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bi-directional Long Short-Term Memory
BRNN	Bidirectional Recurrent Neural Network
CR	Compression ratio
CNN	Convolutional Neural Network
CR	Compression ratio
DL	Deep Learning
DSSM	Deep Structured Semantic Models
DUC	Document Understanding Conference
GRU	Gated Recurrent Unit
IPA	International phonetic writing
KNN	K-nearest neighbor algorithm
LSTM	Long Short Term Memory
DL	Deep learning
MT	Machine translation
NLP	Natural language processing
OOV	Out-of-vocabulary words
RNN	Recurrent neural network
R2N2	Recursive Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RR	Retention Ration
Seq2seq	Sequence-to-sequence model
STDS	Subtopic-driven multi-document summarization model
TAC	Text Analysis Conference

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Text summarization is the technique, which automatically creates an abstract or extractive summary of a text. Text summarization is one of the research works in NLP, which concentrates on providing meaningful summary using various NLP tools and techniques. Abstractive and extractive summarizations are two methods of generating summaries from texts. This study has identified the "Afan Oromo Text Summarization in Deep Learning" as a research topic. The primary purpose of the study is to design a system and implement extractive and abstractive Afan Oromo proclamation text summarization to come up with effective and efficient summarization type as well as to evaluate the extent of the fitness of the algorithms. Hence, 583 articles of 27 Afan Oromo proclamations were used as an input data for the purpose. Accordingly, abstractive text summarization models (Sequence- 2-Sequence decoder with attention) and extractive text summarization models (TextRank) was developed for text summarization of the dataset. Different comparison measures (Rouge-1 and Rouge-2 percentage, count vectorizer, tf-idf vectorizer, and soft-cosine similarity) were implemented to evaluate the text summaries produced. Results of the Rouge-1 and Rouge-2 measurement percentage index were higher for abstractive summarization than that of the extractive one in this case. Besides the algorithms and models used for both summarization methods fit for the Afan Oromo proclamation text summarizations.

Keywords: - Deep Learning, TextRank, RNNs, Attention model, Encoding, Decoding

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the Study

In this era of the age of information, there is swift development in the field of information science. The fast growing means of information technology such as internet and various web pages have opened access for voluminous of data posting formats. The advancement of natural language processing (NLP) has eased the users' understanding of the posted texts and reduced their time of reading the textual information in the form of digital documents. NLP is the field of computer science, artificial intelligence, and linguistics concerned with the interaction between computers and human language [1]. Due to the fact that computers are required to understand what humans have written and produce readable outputs that are user friendly. Hence, NLP is used for parsing or describing, reduction of words and to generate text summary in either extractive or abstractive text summarization methods [2]. In order to attract the reader's appetite in terms of making them understand in precisely shorter time span the need for text summarization has played paramount role. Text summarization is defined as a way to reduce the bulky amount of information into brief and precise form by the process of selecting important information and discarding unwanted and redundant ones. Or the original text can be reduced to shorter readable texts by paraphrasing the large texts without compromising its semantic originality [3]. Automatic text summarization is generated by software which is coherent and contain significant amount of relevant information from the source text. The main target of generating summaries by both human and algorithms is to select, reformulate and create a coherent text containing the most informative segment of the documents [3]. To do so, two types of automatic text summarizations are available, namely, extractive and abstractive text summarization methods [4]. The extractive text summarization method picks up the most important phrases and lines from the documents. It then combines all the vital lines to create the summary. Extractive summarization uses statistical approach for selecting the important sentences or keyword from document [4]. Extracted sentences tend to be longer than average. Conflicting information may not be presented accurately in extractive summarizations [4].

On the other hand, abstractive summarization involves summaries based on deep learning in which it generates a sentence from a semantic representation and then use natural language generation techniques to create a summary that is closer to what a human brain might generate [4]. Such a summary might contain words not explicitly present in the original. It consists of understanding the original text and re-telling it in fewer words. It uses linguistic approach to understand the original text and then generates summary.
Ziqiang, et. al. [5] strongly contend that abstractive summaries are more accurate as compared to the extractive summary but are difficult to generate because it needs deep understanding of the NLP tasks. Abstractive and extractive summarization uses either statistical or linguistics approaches or combination of both to generate summary.

Afan Oromo is one of the natural languages spoken widely in Ethiopia. According to the Ethiopia Population Census Commission 2007 the total population of Oromo People which use the Afan Oromo as their mother tongue is 34.5% of the total population of Ethiopia [6]. It is also the third most widely spoken language in Africa next to Arabic and Hausa languages [7].The language is used as the official working language and medium of instruction in Oromia Regional State. Afan Oromo uses Latin-based alphabets called “Qubee”. Latin based Afan Oromo, “Qubee” alphabet, the writing system. Qubee has 33 characters representing distinct sounds [8]. It has both capital and small letters. Afan Oromo has a considerable amount of glottal stops. An apostrophe, and less commonly a hyphen, is used to represent this sound in writing. Sometimes an H, represents the apostrophe. For a reason to be apparent later, the apostrophe is considered as a distinct symbol (say, as the 27th letter of the alphabet) [8]. In the Afan Oromo writing system, geminated consonants (e.g. ch, dh, ph, etc.) and long vowels (aa, ee, ii, oo, etc.) are represented by double letters. In addition to seven compound symbols Qubee, not all the 26 letters correspond with their English sound representation as shown in the table below.

Table 1.1: The Qubee letters and their phonetic transcription in international phonetic writing (IPA) (Source:[8]Ibrahim Bedhane, 2015)

Qubee		IPA	Qubee		IPA	Qubee		IPA
A	A	/a/	L	/l/	/l/	W	w	/w/
B	B	/b/	M	/m/	/m/	X	t’	/t’/
C	C	/č’/	N	/n/	/n/	Y	y	/y/
D	D	/d/	O	/o/	/o/	Z	z	/z/
E	E	/e/	P	/p/	/p/	CH	ch	/č/
F	F	/f/	Q	/k’/	/k’/	DH	dh	/đ/
G	G	/g/	R	/r/	/r/	NY	ny	/ň/
H	H	/h/	S	/s/	/s/	PH	ph	/p’/
I	I	/i/	T	/t/	/t/	SH	sh	/š/
J	J	/ǧ/	U	/u/	/u/	TS	ts	/ s’/
K	K	/k/	V	/v/	/v/	ZH	ž	/ž/

Furthermore, the Afan Oromo alphabets and its morphology are discussed in detail in chapter three. Besides its various uses, the language is also used as the medium to post-digital texts on web pages. One of the digitally posted texts, the Afan Oromo proclamation texts is the concern of this study. The proclamations are announced in three languages, namely, Afan Oromo, Amharic, and English on “Magalet” Magazine which is one of the popular magazines in Oromia Regional State. In order to make easy access for users, it is necessary to avail the major essence of the proclamations to be easily understood in shorter time span through the natural language processing mechanism called text summarization.

The need for automatic summarization increases as the amount of textual information increases. A lot of information is available on the internet but to sort out the required information is a tedious job. The same is true for Afan Oromo proclamation texts that are periodically posted on web pages of oag.gov.et and/or www.caffeeoromiyaa.org. The need for technologies that can do all the sorting and quickly identify the relevant information on their own therefore plays an important role. Summaries are important when it comes to the process of the huge amount of information. According to [9], summarization has many benefits such as saving time, easing selection and search, improving indexing efficiency, etc. Summaries of the same document can be different from person to person when summarized by individuals manually. This can occur due to different points of interest or due to each individual's understanding. Such a manual summarization exposes the summarized texts to bias and is also time-consuming. To avoid bias, repetitions, and inconsistencies technology-based text summarization were put into effect.

Prior sample summarization related research works done on the Afan Oromo texts are 'Afan Oromo News Text Summarizer', by Girma Debele (2012); 'Afan Oromo Text Summarization Using Word Embedding', by Lamesa Teshome (2020); 'Afan Oromo Text Summarization Using Sentence Scoring' by Gammachiis Temesgen (2021); 'Afan Oromo Automatic News Text Summarizer Based on Sentence Selection Function', by Fisseha Berhanu (2013), 'Information Extraction Model From Afan Oromo News Texts', by Sisay Abera (2020) and others [10] , [11], [12], [13] and [14].

As far as the researcher's knowledge is concerned there are no researches that have done on Afan Oromo Proclamation texts. Most of the text summarization researches were done on news texts. Here comes the need for the summarization of extractive and abstractive text summarizations for the Afan Oromo proclamation texts in order to evaluate their results. It was also found importance to compare the experiment result of the current work with prior text summarization research works done on Afan Oromo texts.

1.2. Statement of the Problem

It is obvious that from the twenty first century onwards it was considered to be the era of information due to the explosion of large amounts of information. Voluminous information segments of various types are posted digitally on web pages. Currently, digital information are widely spreading across the globe by the use of news script media advanced information technology [9]. The rapid growth of competitive technology has made the lifestyle of people difficult to the survival if not using time properly. Hence, there is a need to spare the scarce time to distribute properly across their business. Without the access to information it is hardily to lead better life. But the availed information is bulky and voluminous that may consume the scares time of the users. For the information to be easily accessed it needs to be short, concise, easily understandable and has to contain the salient important issues of the total text document. To facilitate the process of shortening and making user friendly text information on internet, one of the natural language processing (NLP) approaches called text summarization was used for summarization of different digital texts in different languages. There are numerous text summarizations done on English text documents due to the pre-prepared suitable corpus of the language [15] and [16]. There are also fairy good text summarization research works done on Amharic [17]. To the contrary, there are very few text summarization researches conducted on Afan Oromo text documents due to the lack of pre-prepared corpus as the result of the unique nature of the language [10], [11], [12] and [13]. To alleviate this problem this study targeted to deal with Afan Oromo text summarization in deep learning with special reference to Afan Oromo proclamation texts. It also strives to see the current summary results with that of the prior related studies done on Afan Oromo. Accordingly, this study attempts to address the following basic research questions.

- What system design is developed to implement extractive and abstractive Afan Oromo proclamation texts?
- What evaluation metric is used to evaluate the outcomes of the Afan Oromo proclamation texts?
- What results are obtained from the evaluation of the texts' summaries?
- What results are obtained when the current research result is seen against the prior related research works done on Afan Oromo documents text?

1.3. Objectives of the Study

The study has objectives to accomplish. There are general and specific objectives that are outlined here.

1.3.1. General Objective

The general objective of the study is to Summarize Afan Oromo proclamation test documents and to compare the results of this study with the results of the prior related research works done on Afan Oromo texts summarizations.

1.3.2. Specific Objectives

Based on the aforementioned general objective the following specific objectives are presented as follows.

- To develop system design and architecture that could be used for effective extractive and abstractive Afan Oromo proclamation texts summarization.
- To implement abstractive and extractive text summarization by using the developed system designs and architecture on the Afan Oromo proclamation texts
- To evaluate the qualities of the extractive and abstractive text summarization results by using appropriate evaluation metrics.
- To compare the Afan Oromo proclamation text summarization methods results with prior research results done on Afan Oromo texts.

1.4. Significance of the study

As was aforementioned, Afan Oromo is widely spoken in Ethiopia and it is also spoken in some countries around Ethiopia (such as in Kenya, Uganda, and Somalia). Afan Oromo is one of the languages used to produce proclamation document texts occasionally. This indicates that when numbers of proclamations texts were posted digitally it needs to give ease access for the users. Nevertheless, reading the documents as it is will take longer time and the main ideas might not be captured by the users. Hence, distilling the main idea from the texts through summarization is one of the important steps. Besides it is also important to use deep learning approach to make the hidden ideas clear and avail concise and clear summaries of idea for the users to make them understand efficiently. Therefore, this study becomes important due to the fact that automatic summarization of the proclamation texts in deep learning approach will enhance the deep understanding of the texts besides saving the time of users.

The need for automatic text summarization increases as the amount of textual information increases. A lot of information is available on internet but to sort out the required information is tedious job. Text summarization is the process of extracting the important information and gives the overall idea of the entire document. It is a tedious task for human beings to generate an abstract summary manually since it requires a rigorous analysis of the document. In order to ease human efforts and to reduce time automatic summarization techniques prove to be very useful. Automatic text summarization is a technique which can automatically generate the desired and relevant information from a huge amount of information. The goal of automatic summarization is to form a shorter version of the source document by preserving its original meaning and information content. In this context, comparing the extractive and abstractive text types is of paramount importance for making effective and efficient text summarization. This will ease the user's time constraints and understanding of the document. Hence, text summarization aims at composing a concise version of the original text, retaining its salient information since manual text summarization is a demanding time expensive and generally laborious task. In addition, automatic text summarizations gaining increasing popularity and therefore, constitutes a strong motivation for further research.

1.5. Scope and Limitations

It is of paramount importance to indicate and be guided by the scope, limitations and delimitations of study to focus only on the intended issues. Therefore, the following scope, limitations and delimitations were presented.

1.5.1. Scope of the study

Generally, the scope of the study had referred to the parameter under which the study had operated. Thus, this study dealt with text summarization on Afan Oromo. As text summarization is a wide concept in general and that of Afan Oromo text summarization in particular. Specifically, among the other documents, the study was concerned with the summarization of Afan Oromo proclamation texts. That is, the scope of this study was the extractive and the abstractive summarization of 583 articles of 27 Afan Oromo Proclamations texts using deep learning.

1.5.2. Limitations of the study

Limitations are matters and occurrences that arise in the study which are out of the researcher's control. Five hundred eighty three articles of 27 Afan Oromo Proclamations were only used for the analysis. The reasons behind the selecting these proclamations is that the researcher is obliged to limit the numbers due limited numbers of the digitally posted proclamations of Afan Oromo.

CHAPTER TWO

1. REVIEW OF RELATED LITERATURE

Relevant reviews of literature and research works on Afan Oromo text summarization is presented this chapter.

2.1. Text Summarization

At present time, due to the availability of voluminous digitally posted textual data there is a need to present them in user friendly style [18].Text summarization is presenting the precise, condensed and meaningful short texts that represent the necessary contents of the original document.

There are some ambiguities that that are prevalent in automatic text summarization that is linked directly with text understanding that results in some challenges. The challenges might come due disparities in text formats, expressions and editions [18]. Natural language processing is one of the means by which researchers approach problems in text summarization by focusing on the fundamental issue to identify the text focus [19].

The retrieval of important information from long texts within shorter time, easy and rapid loading of the salient information, and resolution of the problem associated with the criteria needed required the application of automatic text summarization [20].

Among the others, one of the main reasons they explained by Chen et al. [20] was that the progress and advancement of technology use of automatic text summarization methods had enormous results. However, the methods need to be reviewed and evaluated from time to time with the advancement of technology.

Text summarizations could be divided based on categories such as its function, genre, context, summary type and document numbers [21]. From the various categories one is extractive and abstractive text summarizations is one of the categories [15]. Gu J. et.al [15] explain further that extractive summarization selects some parts of the original text based on the computation of either the statistical features or linguistic features that focus on the salient sentences, phrases and parts of the sentence.

The other category is abstractive text summarization. It generates new phrases by paraphrasing or restating the original text which makes the task of computer the tough one [16]. There are also other factors that make abstractive text summarization more difficult. One of the factors indicated is that scientific knowledge and semantic analysis are needed for the abstractive summarizations [16]. What makes abstractive summarization to be better is that it more meaningful and it nearly resembles the summary generated by human beings [22]. Acceptable and meaningful summaries of both abstractive and extractive need to have logical coherence, minimum repetition and major concepts in the original text [23].

In order to transmit representative, meaningful and important information the generated summaries must be of required standard [16, 24]. Khan, et. al. [25] stated that different schemas such as lead, body phrases, ontology, tree, template and rule-based are encoded by structured abstractive summarization approach. The information representation of the summarized document focuses more on meaning and it is based on semantic approach. According to Jafar [26], the semantic based approaches include the multimodal method, information item method and the semantic graph based method.

2.1.1. Types of summaries and their properties

Text summarizations are classified into five types [26] [27]. They are indicative versus informative, extractive versus abstractive, generalized versus query-based, single versus multi document summarization and shallow versus deep summarizations.

2.1.1.1. Indicative versus Informative

According to Indejeet [29], when the reader focuses on what the information is all about he uses indicative summary. And also when the wants a representing information about the text he uses informative summary. Characteristics of the main document such as length and the writing style can be gained from the indicative one while representative facts about the information in the document be found by the informative summarization type.

2.1.1.2. Extractive versus Abstractive

When the summary is based on the important original sentences or phrases taken directly as they appear, it is called extractive summarization. On the other hand, when the summary is written by sentences and phrases that are not in the original document but represent the meaning and concept of the original one in precise and shorter summary output it is called abstractive text summarization [30]. Pachantouris, G. [30] stated the following inherent problems in both types of summaries.

The problems with extractive summarization methods are:

- There are times when generated summaries by extraction become more than the amount the average sentences. This happens when the unwanted sentences and phrases are included in the summary.
- The other problem is that when the summary fails to capture the important sentences or phrases that convey the proper idea.
- And also there are times when there is no representation of contradictory information

Also there are problems that might be inherent in abstractive text are [28]:

- When the user chooses extractive summarization instead of the abstractive one due to the reason of preferring the summary line by line that is similar with the original text.
- There are times when incoherence occurs at the border of two sentences.
- The complexity of the human language and the absence pre-prepared lexicon might lead to poor the summary outputs.

2.1.1.3. Generalized versus Query-based Summarization

Generic and query-based is another type of classification of text summarization based on the preference of the user. The generic summary assists the reader to quickly determine about the document by just only reading part of it [32].

To the contrary, the goal of query-based summarization is to summarize the information in the input document(s) which is pertinent to the specific user query [33]. In this context, the summarizer tries to find information relevant to the query or in some cases, may indicate how much information in the document relates to the query. According to Ani [33], much of the work to date has been in the context of generic summarization, nevertheless, in order to generate important and useful summary, query-based summary is recommended to be used by the summarizer.

2.1.1.4. Single versus Multi Document Summarization

Single and multi-document summary are the other bases on which text summarizations are classified. Single document summarization is systems produced a summary of one document. Examples of single document summaries are scientific article, court decision document, lecture, broad casted single topic media news [34].

The cluster of news articles on the same event to produce online browsing pages of the current events could be multi-document summarization. Such large amount of repetition on the web is boring and time consuming. Hence, there is a need for multi-document summarization to avail brief and meaningful summary to the user [35].

2.1.1.5. Shallow versus Deeper Summarization

The level in the linguistic space becomes a base for another method of summary classification called shallow versus deeper summarization [36]. The shallow approach focuses on the grammatical level of representation during text summarization that does not refer to the semantic aspect of the original text. To the contrary, the deeper text summarization deals with semantic level of representation of the original text based on linguistic processing.

2.1.2. Stages of Text Summarization

According to Spark [35], quality summarization demands a clear thorough analysis of context factors. He [35] distinguished three classes of context factors, namely, input factor, purpose factor and output factor. Their brief description is presented as follows.

- The input factor determines the way the structure of the text to be summarized such as text form, subject type and unit.
- The purpose factor indicates for whom the summary is prepared ad what the summary is for. It is concerned with audience, and use.
- The material to be summarized as well as the format and the style of the summarization are the output factors.

2.2. Deep Learning

Deep learning (DL) is part of a machine learning that uses algorithms and to produce high level abstraction in data [36]. Dl algorithms develop and model layered hierarchical learning architecture that resembles the function of the human brain which automatically extracts features and abstractions from the data under summarization process [37] [38] [39]. Regarding large amount of unsupervised data, algorithms are useful to learn the data representation in a greedy layer-wise method.

2.2.1. Importance of Deep Learning

One of the major uses of deep learning is that it minimized or facilitated the difficult and time consuming voluminous textual documents summarization [40]. It is also used in varieties different professions such as in natural language processing, speech recognition, medical informatics, and other application areas. It also offered a unified set of tools for tracking diverse problems [41].

2.2.2. Application of Deep learning

There are seven types of DL applications, namely, Automatic Speech Recognition, Image Recognition, Natural Language Processing, Drug Discovery and Toxicology. Customer Relationship Management, Recommendation Systems and Bioinformatics [41]. One of the basic uses deep learning techniques is to discover the hidden structures and features at different levels of abstractions which are useful for the any language processing tasks.

2.2.3. Deep Learning Models

Recurrent neural network (RNN), Bidirectional RNN (BRNN), Attention Mechanisms (AM) , Long Short-term Memory (LSTM), Gated Recurrent Unit (GRU), and Sequence to Sequence Models (Seq2seq) are the six models of deep learning [16]. Each are briefly discussed as follows.

2.2.3.1. Recurrent Neural Networks based Models

As a class of neural network, Recurrent Neural Network (RNN) [42] has outperformed in modeling sequential data. Sequential relations and syntactic/semantic information from word sequences are captured by RNN. The neurons which are inside RNN models are connected through hidden layers. The inputs of RNN are from word or sentence embedding as well as from the outputs of the previous hidden state this makes RNN a powerful model. Thus, a large number of RNN-variants which is the most prevalent ones are Long Short-Term Memory (LSTM) [43], Gated Recurrent Unit (GRU) and Bi-directional Long Short-Term Memory (Bi-LSTM) [44].

Li et al. [45] developed a RNN-based framework to estimate the salience information from documents in an unsupervised manner. This framework is employed to extract salience information vectors from the input sentences. Also the process of the cascaded attention retains the most relevant embedding to reconstruct the original input sentence vectors and trivial information in the output vectors are penalized during construction process of the model [45].

The idea of multi-document summarization assumes the set of documents belongs to the same topic but within the documents set, documents may belong to the same/different subtopics and the importance of sentences vary when it comes to different subtopics. Extracting sentence, document and topic embedding from the original documents were undertaken by the use of hierarchical RNN structure [46].

2.2.3.2. Bidirectional RNN

Indicated that Bidirectional RNN consists of the forward RNNs and backward RNNs were indicated by Zheng [46]. A sequence of hidden states is generated by forward RNN after reading the input sequence from left to right. On the other hand, a sequence of hidden states is generated by the backward RNNs after reading the input sequence from right to left. The concatenation of the forward and backward RNNs is the representation of the input sequence [47]. Therefore, the representation of each word depends on the representation of the preceding (past) and following (future) words. In this case, the context will contain the words to the left and the words to the right of the current word [48] [60]. Schuster, et. al. [48] concluded that using bidirectional RNN enhanced the performance.

2.2.3.3. Attention Mechanism

A basic encoder-decoder architecture cannot consider all the elements of long input which makes it fail since the size of encoding is fixed for the input string [49]. One of the important uses of attention mechanism is to remember the input that has significant impact on the summary [50]. To calculate the weight between the output word and every input word attention mechanism is used. The calculated weight adds to one.

2.2.3.4. Long Short Term Memory (LSTM)

Input/read, memory/update, forget, and output gates are the repeating unit of LSTM architecture [51] but the chaining structure is the same as that of an RNN. The four gates that share information with each other are: Input Gate, Forget Gate, Memory Gate and Output Gate, the Gates facilitate information flow for a long period of time [51] each is discussed by Lopyrev, K. [51] as follows.

- The ***Input Gate*** is a vector that is randomly initialized following subsequent steps in which the input of the current step is the output of the previous step. The input is subject to element-wise multiplication with the output of the forget gate in all the cases in which the multiplication result is added to the current memory gate output.
- ***A Forget Gate*** is a neural network that has one layer and a sigmoid activation function. The value of will determine If the information of the previous state is determined by the sigmoid function whether to be forgotten (when the sigmoid value is 0) or remembered (when the sigmoid value is 1). The output of the previous block, the input vector, the remembered information from the previous block, and the bias are the four inputs for the forget gate.
- ***Memory Gate***. The effect of the remembered information on the new information is controlled by the memory gate which consists of two neural networks. The first network has different bias that makes it distinct. The second neural network has a tanh activation function which is utilized to generate the new information.
- ***Output Gate***. The output gates control the amount of new information that is forwarded to the next LSTM unit is controlled by the output gates with a sigmoid activation function that considers the input vector, the previous hidden state, the new information, and the bias as input. The output of the sigmoid function is multiplied by the tanh of the new information to produce the output of the current block.

2.2.3.5. Gated Recurrent Unit (GRU)

Another name for simplified LSTM is GRU that has two gates: a reset gate and update gate and it has no explicit memory. When all reset gate elements approach zero, the previous hidden state of information is forgotten as the result the only input vector affects the candidate of the hidden state. This is the case when the update gate acts as a forget gate. Abstractive text summarization uses both LSTM and GRU commonly to provide extra control.

A GRU is a simplified LSTM with two gates is called, a reset gate and an update gate, and there is no explicit memory. The previous hidden state information is forgotten when all the reset gate elements approach zero; then, only the input vector affects the candidate hidden state. In this case, the update gate acts as a forget gate. LSTM and GRU are commonly employed for abstractive summarization since LSTM has a memory unit that provides extra control; however,

2.2.3.6. Sequence to Sequence Models

In the framework of sequence-to-sequence models, a very relevant model to the text summarization task is the attentional RNN encoder-decoder model that is proposed by Cao, et.al [55], which has produced state-of-the-art performance in machine translation (MT). Since MT is also a task that maps one word-sequence to another, the attentional RNN encoder-decoder is a natural candidate for summarization too.

2.2.4. Evaluation of Text Summarization

Text summary results whether human or machine produced require evaluation. But evaluation of text summary had been a difficult task due to the reason that the prevalent use of various metrics and the lack of a standard evaluation metric. In this sub section the automatic evaluation issues are presented.

2.2.4.1. Human Evaluation

Human made evaluation is one of the simplest ways of evaluation to obtain quality summary.

For instance, in DUC, the judges would evaluate the coverage of the summary, i.e. how much the candidate summary covered the original given input. In more recent paradigms, in particular TAC, query-based summaries have been created. Then judges evaluate to what extent a summary answers the given query. The factors that human experts must consider when giving scores to each candidate summary are grammar, non-redundancy, integration of most important pieces of information, structure and coherence [52].

2.2.4.2. Automatic Evaluation Methods

Various automatic text summarization metrics were used as of early 2000s of which Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is the one. Currently, ROUGE is the most widely used metric for automatic evaluation [52]. ROUGE as one of the most essential evaluation metric it consists of collection of evaluation indicators for natural language processing.

Machine translation, single document and multi-document summarizations are the components of natural language processing that are to be evaluated by ROUGE. According to Lin, C. [52] ROUGE as an evaluation metric has many variants by which it measures the output abstracts in multiple ways through comparing them. Currently, ROUGE N and ROUGE L are the commonly used evaluation metric for automated text summarization.

ROUGE-N [52] is an evaluation metric that stands for ROUGE with N-gram Co-Occurrence Statistics. ROUGE N measures an n-gram recall between reference summaries and their corresponding candidate summaries. Formally, ROUGE-N can be calculated as:

$$ROUGE - N = \frac{\sum_{S\in\{Ref\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S\in\{Ref\}} \sum_{gram_n \in S} Count(gram_n)},$$

Figure 2.2.4.2: ROUGE-N formula

Here there is an explanation given by Lin [52] about this figure, that is, $Re\ f$ is the reference summaries and n represents n-gram. $Countmatc$□ $(gramn)$ represents the maximum number of n-grams in the reference summaries and the corresponding candidates.

The numerator of ROUGE-N is the number of n-grams owned by both the reference summaries and the automatically generated ones, while the denominator is the total number of n-grams occurring in the golden summary. The denominator could be set to the number of candidate summary n-grams as well to measure the precision. However, ROUGE-N mainly focuses on quantifying recall, so precision is not calculated here.

The special cases of ROUGE-N that are chosen as the best practices are ROUGE-1 and ROUGE- 2. The unigram is represented by number 1 and the bigram is represented by number 2. Most of the research works adopts and uses 1, ROUGE-2 or ROUGE-N. Hence, ROUGE-1, ROUGE-2 and ROUGE-L are used in this research for the evaluation purposes. ROUGE evaluation metric produce recall, precision and F-score values.

How many of the sentences in the reference summary that are present in the generated summary is measured by sentence recall. Precision measures the extent to which the machine generated summary exist in the manually generated summary. Precision and Recall are standard measures for Information Retrieval and are often combined in a so-called F-score.

Precision: The ratio of sentence that exist both in manually generated summary and machine generated summary per machine generated summary.

$$\text{Precision}= \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

Recall: The ratio of sentence that exists in both manual generated summary and machine generated summary per manual generated summary.

$$\text{Recall}= \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

F- Measure: This is harmonic mean of precision and recall.

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

2.3. Review of Prior Researches Works

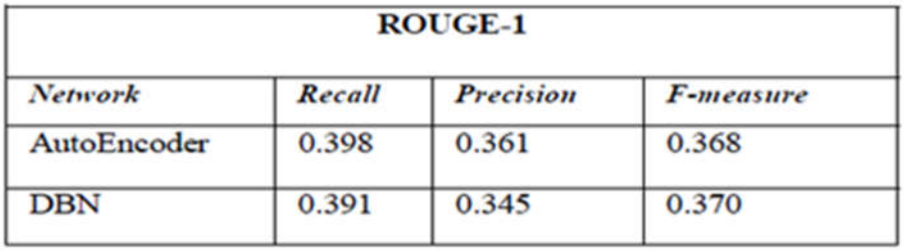
Various research works were done on extractive and abstractive text summarizations. In this section the summarization methods, methodologies, models and techniques used as well as the results obtained are discussed. The rationale behind referring to these research works is to compare the result of the current research against the results of some the prior text summarization research works done on other languages such as on English and Amharic as well as Afan Oromo.

2.3.1. Text summarization research on English

2.3.1.1. Extractive Text Summarization (English)

Afsaneh Rezaei, et. al. [53] conducted text summarization research on “Multi- Document Extractive Summarization via Deep Learning Approach”. The main goal of the study is to use deep learning (in the form of deep auto-encoder neural network and deep belief network (DBN) to generate a multi-document extractive summarization. The major objective of this research is to use auto-encoder neural network and deep belief network separately for scoring sentences in a document to compare their performances. These systems are implemented on Document Under Conference (DUC) 2007 dataset and evaluated by using ROUGE tool. The abovementioned systems were tested on Document Under Conference (DUC) 2007 dataset and evaluated using ROUGE-1 and ROUGE-2 criteria.

Table 2.3.3.1: Average Results of ROUGE 1 for Similar Unigram

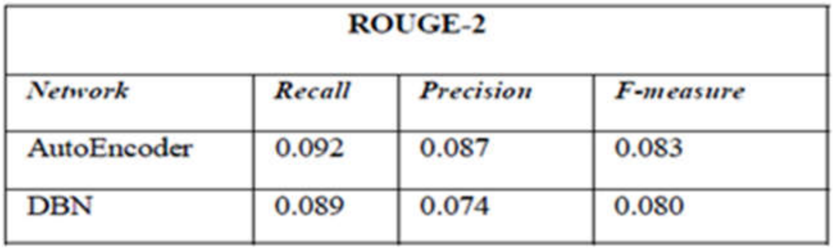


ROUGE-1			
Network	*Recall*	*Precision*	*F-measure*
AutoEncoder	0.398	0.361	0.368
DBN	0.391	0.345	0.370

Source: Afsaneh Rezaei, et. al. [1]

In each of the above methods, the three important criteria of the Recall, Precision, and F-measure were also calculated. Table 2.3.1.1 of the study shows the evaluation results of summaries generated by the two proposed networks using ROUGE tool. Here, the three criteria of recall, precision, and f-measure were calculated based on similar unigrams. As can be seen, the F-measure from DBN (F-measure=0.370) shows a higher number with negligible difference. But the results from the two networks indicated that the auto-encoder network generates more reliable summaries (Precision=0.361) than DBN does (Precision=0.345).

Table 2.3.1.2: Average Results of ROUGE 2 for Similar Bigram

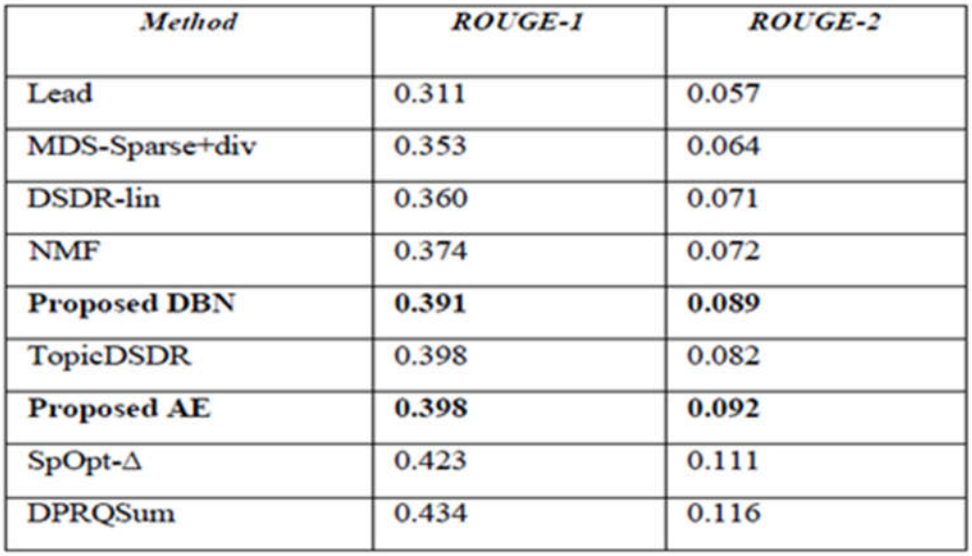


ROUGE-2			
Network	*Recall*	*Precision*	*F-measure*
AutoEncoder	0.092	0.087	0.083
DBN	0.089	0.074	0.080

Source: Afsaneh Rezaei, et. al. [1]

Also in Table 2.3.1.2, the evaluation results of similar bigrams are shown. The comparison between the two networks shows that in general, the auto-encoder network has a better performance (Recall= 0.092, Precision= 0.087 and F-measure= 0.083) than DBN (Recall= 0.089, Precision= 0.074 and F-measure= 0.080).

Table 2.3.1.2: Performance Comparison of Different Methods Based on ROUBE 1 and ROUGE 2



Method	*ROUGE-1*	*ROUGE-2*
Lead	0.311	0.057
MDS-Sparse+div	0.353	0.064
DSDR-lin	0.360	0.071
NMF	0.374	0.072
Proposed DBN	**0.391**	**0.089**
TopicDSDR	0.398	0.082
Proposed AE	**0.398**	**0.092**
SpOpt-Δ	0.423	0.111
DPRQSum	0.434	0.116

Source: Afsaneh Rezaei, et. al. [1]

ROUGE-1 and ROUGE-2 values were compared for various methods. Table 3, shows that the proposed DBN methods (0.089) for ROUGE-2 showed better scores when compared to other methods, and ROUGE-1 results from summaries of the 'Proposed AE' methods (0.398) had acceptable values relative to other methods.

Based on the results of the study they concluded that the auto-encoder network has generally better performance than DBN. Compared to other systems, DBN with mean ROUGE-2 equal to 0.089 and an auto-encoder equal to 0.088 showed better results. Moreover, compared to the currently available methods, the values obtained from these systems for ROUGE-1 are acceptable and reliable.

2.3.1.2. Abstractive Text Summarization (English)

Mohammad Masum, et, al. [54] done a text summarization research on, "Abstractive Method of Text Summarization wit Sequence to Sequence RNN". Amazon fine food reviews dataset was used for the purpose of summarization. Text descriptions were taken as inputs that generated simple summary of the review descriptions as output. The algorithm RNN with LSTM's in encoding layer and attention model in decoding layer were used. Sequence to sequence model was applied to generate a short summary of food description. In the process of the study, the researchers faced with some challenges. The challenges working with the abstractive text summarizer are text processing, vocabulary counting, missing word counting, word embedding, the efficiency of the model or reduce value of loss and response machine fluent summary. The main goal of this study is to increase the efficiency and reduce train loss of sequence to sequence model for making better abstractive text summarizer. The experiment results showed that the training loss was successfully reduced to the value of 0.36 and better short abstractive text summarizations were created from English to English texts.

2.3.2. Text summarization research done on Amharic

Abaynew Guadie, et.al [55] conducted text summarization research entitled 'Amharic Text Summarization for News Item Posted on Social Media. The study addressed the problem of the summarizing user posts of interest in a human reader by extracting the most representative posts of the irrelevant Amharic tweet for the post news that could be summarizing without duplicate posted Amharic text documents. The researchers proposed three main approaches. First, the similarity between each posted document within the two pair of sentences is calculated. Second, by using Kmeans algorithm clustering based on the similarity results of the documents were grouped. Third, summarizing the clustered posted document individually using TF-IDF algorithms that involve finding statistical ways for the frequent terms to rank the documents. Three main components based approach was used to undertake the work step by step. First, calculate the similarity between each posted document within the two pair of sentences. Second, clustering based on the similarity results of the documents to group them by using Kmeans algorithm. Third, summarizing the clustered posted document individually using TF-IDF algorithms that involve finding statistical ways for the frequent terms to rank the documents. The researcher applied the extractive summarization technique to extract the sentences with highest ranked sentences in the posted documents to form the summaries and the size of the summary can be identified by the user. The following experiment results were obtained. The result of experiment 1 showed 87.07%, which is the highest F-measure score for extraction rate of 30%. The result of the second experiment was F-measure score of 84% for extraction rate of 30%. In the third experiment the highest F-measure score is 91.37% for extraction rate at 30%. The fourth experiments the highest F-measure score is 93.52% for extraction rate at 30% to generate the summary post texts. If the system to generate the size of summary is increased, the extraction rate also increased in posted texts. They conclude that the evaluation system shown that a very good results to summaries the posted texts on social media.

2.3.3. Text summarization researches done on Afan Oromo

The following three prior Afan Oromo text summarization research works were reviewed.

Lamesa Tashoma [10] conducted a research titled “Afan Oromo Text Summarizer Using Word Embbeding”. The main focus of the study was to develop a generic automatic text summarizer for Afan Oromo text based on word embedding. The language specific elements such as stop words and stemmer or lexicons were used to develop the summarization. To select important sentences for the summary from the text document graph based PageRank algorithm was used. Cosine similarity was used to measure similarity between the sentences. The stop word list, suffix and other language specific lexicons were collected from previous Afan Oromo summarization research works. The system performance has been evaluated based on three experimental scenarios and evaluation. The achieved results were precision = 0.527, recall = 0.422 and F-measure =

0.468 by using the data we gathered. And also the overall performance of the summarizer outperformed by precision = 0.648, recall = 0.626 and F-measure = 0.058 when compared with the previous works by using the same data used in their work.

Fiseha Berhanu [11] on its part conducted a research titled “Afan Oromo Automatic News Summarizer Based on Sentence Selection Function”. The focus of the research work was to developing Afan Oromo news text summarizer. This was done by systematically integrating the features of the language such as sentence position, key word frequency, cue phrase, sentence length handler, occurrence of the numbers and events like time, date and month. Abbreviations, synonyms, stop words, suffixes, numbers and names in general and names of time, date and months were collected from both the secondary and primary sources to aid the development of the system. Besides, 350 English cue words were collected and translated to Afan Oromo cue phrases.

The result of subjective evaluation is 88% informativeness, 75% referential integrity and non-redundancy, and 68% coherence. Because of the added features, different techniques and experiment applied to this work the system gave 87.47% and outperform by 26.95% than the previous work

Gammachis Temesgen [12] done a text summarization research work under the title “ Afan Oromo News Text summarization Using Sentence Scoring Method”, the work was undertaken by using features like thematic words, word frequency, title words, term weight, cue phrases, name of numbers, and sentence position. Ten news topics that were selected out of thirty were used for the extractive summarization purpose. The preparation of manual summary was executed by three speakers of the language. Using the python programing NLTK system was developed. As the result of feeding each individual word to the system, the system generated three extraction rate, namely, 2o%, 30%, and 40%. The results indicated 72%, 82%, and 84% which is the extraction rate performance. The results of the coherence of the evaluation summary showed 62%, 66% and 72%. The informativeness of the rate of the performance of the summary showed 74%, 78%, and 86%. The average of the evaluation of the three metrics recall, precision, and F-score showed 86.1%of performance by the system.

Critics

The following shortcomings were observed by the researcher in the reviewed prior research works.

The English works

The Extractive one (For English)**:-**

- The summary work was done only on short text.
- The maximum output of long text provides incorrect summary.
- The hardware configuration took longer time during the process

The Extractive one (For English):-

- There were challenges working with the abstractive text summarize, such as text processing, vocabulary counting, missing word counting, word embedding, the efficiency of the model or reduce value of loss and response machine fluent summary.

The Amharic works

- There is no pre-prepared uniform Amharic lexicon just like the developed lexicon of English.
- Only the F-score measure values are produced.
- There are no recall values and precision values calculated which makes the summary output results vague. When the recall is not calculated the extent of the system summary captured from the reference summary is not known. Similarly, when the precision value is not calculated the extent of the relevance of the system summary cannot be indicated.

The Afan Oromo works

- There is no pre-prepared uniform Afan Oromo lexicon just like the developed lexicon of English
- Almost all the text summarization research works were done on extractive text summarization which excluded the abstractive text summarization.

- Though most of the research works were done on Afan Oromo news texts is considered a limitation, however the use of different models and algorithms on the extractive summarization is one of the encouraging tasks.

CHAPTER THREE

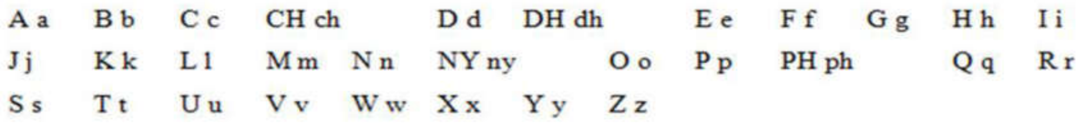
3. THE AFAN OROMO ALPHABETS AND MORPHOLOGY

In this chapter, Afan Oromo Alphabet and writing system, punctuation marks and usage, Afan Oromo morphology, Afan Oromo word, and sentence boundaries are discussed.

3.1. The Afan Oromo Alphabets

Afan Oromo is a phonetic language, which means that it is spoken in the way it is written. The writing system of the language is straightforward which is designed based on the Latin script. Unlike English or other Latin-based languages, there are no skipped or unpronounced sounds/alphabets in the language. Every alphabet is to be pronounced in a clear short/quick or long/stretched sounds. In a word where consonant is doubled the sounds are more emphasized. Besides, in a word where the vowels are doubled the sounds are stretched or elongated [56].

Afan Oromo has the same vowels and consonants as English. Afan Oromo vowels are represented by the five basic letters such as **a, e, i, o, u**. Consonants, on the other hand, do not differ greatly from English, but there are few special combinations such as “**ch**” and “**sh**” (same sound as English), “**dh**” in Afan Oromo is like an English "**d**" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afan Oromo consonant is “**ph**” made when with a smack of the lips toward the outside “**ny**” closely resembles the English sound of “**gn**”. We commonly use these few special combination letters to form words. For instance, **ch** used in **kofalchiisaa** ‘making laugh’, **sh** used in **shamarree** ‘girl’, **dh** used in **dhadhaa** ‘butter’, **ph** used in **buuphaa** ‘egg’, and **ny** used in **nyaata** ‘food’. In general, Afan Oromo has 31 letters (26 consonants including special combinations and 5 vowels) called “**Qubee**” [57]. All the alphabets of Afan Oromo are presented in Figure



A a	B b	C c	CH ch		D d	DH dh		E e	F f	G g	H h	I i
J j	K k	L l	M m	N n	NY ny		O o	P p	PH ph		Q q	R r
S s	T t	U u	V v	W w	X x	Y y	Z z					

Figure 3.1: Afan Oromo Alphabets/ Qubee Afan Oromo

In general, all letters in the English language are also in Afan Oromo except the way it is written.

Words

The word is the smallest unit of a language. There are different methods for separating words from each other. However, most of the world languages including English use the blank character (space) to show the end of a word. Some long words are being cut in written form (abbreviation), with the symbols "/", ".", and therefore this symbol should not determine a word boundary. The usual parenthesis, brackets, quotes, all kinds of marks, are being used to show a word boundary in Afan Oromo [58].

Sentence

Afan Oromo sentence is terminated like English and other languages that follow the Latin writing system [58]. That means, the full stop (.) in a statement, the question mark (?) in interrogative and the exclamation mark (!) in command and exclamatory sentences, mark the end of a sentence and the comma (,) which separates listing in a sentence and the semicolon is to mark a break that is stronger than a comma but not as final as a full stop balance.

3.2. Afan Oromo Morphology

Morphology is a branch of linguistics that studies and describes how words are formed in a language [59]. There are two types of morphology: inflectional and derivational. Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like a person, gender, number, tense, case, and mode. Inflectional changes do not result in changes in parts of speech. On the other hand, derivational morphology deals with those changes that result in changing classes of words 28 (changes in the part of speech). For instance, a noun or an adjective may be derived from a verb.

3.2.1. Types of Morphemes in Afan Oromo

A morpheme is the smallest semantically meaningful unit in a language. A morpheme is not identical to a word, and the principal difference between the two is that a morpheme may or may not stand alone, whereas a word, by definition, is a freestanding unit of meaning. Every word comprises one or more morphemes. In Afan Oromo, there are two categories of morphemes: free and bound morphemes. A free morpheme can stand as word on its own whereas bound morpheme does not occur as a word on its own [57, 58]. In Afan Oromo roots (stems) are bound as they cannot occur on their own. Example: "**dhug**-" (drink) and "**beek-**" (know), which are pronounceable only when other completing affixes are added to them [80].

Similarly, an affix is also a bounded morpheme that cannot occur independently. It is attached in some manner to the root, which serves as a base. These affixes are of three types – prefix, suffix, and infix. The first and the second types of affixes occur at the beginning and at the end of a root respectively in creating a word whereas the infix occurs in between characters of the word. In **dhugaatii** 'drink', for instance, **-aatii** is a suffix and **dhug**- is a stem. Moreover, an infix is a morpheme that is inserted within morpheme. In the work of [59] it is discovered that Afan Oromo does not have infixes like English.

There are many ways of word formation in Afan Oromo. These morphological analyses of the language are organized into six categories [59]. The categories are nouns, verbs, adjectives, adverbs, functional words, and conjunctions. Almost all Afan Oromo nouns in a given text have person, number, gender, and possession markers which are concatenated and affixed to a stem or singular noun form. Afan Oromo verbs are also highly inflected for gender, person, number, and tenses. Adjectives in Afan Oromo are also inflected for gender and number. Moreover, adverbs can be categorized into adverbs of time, adverb of place, and adverb of how some of the adverbs are affixed.

Furthermore, functional words can be classified as prepositions; postpositions, and articles markers which are often indicated through affixes in Afan Oromo. Lastly, conjunctions can be separate words (subordinating or coordinating), and some of them are affixed. Since Afan Oromo is morphologically very productive, derivation, reduplication, and compounding are 29 also common in the language. The following is detail descriptions and examples of the word-formation process of Afan Oromo based on the works of [57, 60].

3.2.1.1. Afan Oromo Nouns

I. Gender

Gender is one category of nouns, pronouns, and adjectives into masculine and feminine and some language neuter based on whether a noun is considered as male, female, or without sex respectively.

Gender is of two types: natural and grammatical.

Natural gender refers to the natural sex of animate things.

Example

Abbaa father -**Haadha** mother

Dhirsa husband -**Niitii** wife

Most nouns are not marked by gender affixes. Only a limited group of nouns differ by using different suffixes for the masculine and the feminine form. Grammatically the language uses -**ssa** for masculine and -**ttii** for the feminine [60].

Obboleessa brother - **obboleettii** sister

Ogeessa expert (male) - **Ogeettii** expert (female)

Natural female gender corresponds to the grammatical feminine gender. The sun, moon, stars, and other astronomic bodies are usually feminine. In some Afan Oromo dialects, geographical terms such as names of towns, countries, rivers, etc. are feminine, in other dialects such terms are treated as masculine nouns. It is due to this fact that there are different subject forms for the noun **biyya** 'country' [58].

Example: **biyyi** (male) or **biitti** (female).

There are also suffixes like -**a**, -**e** that indicate a present and past form of masculine markers respectively. -**ti** and -**tii** for present feminine marker and -**te** past tense marker, -**du** for making adjective form [59].

Biiftuun **baate** 'the sun rose'. 30 The word **baate** takes -**te** to show feminine gender. We can see that -**tii** can also show feminine gender in the following statement.

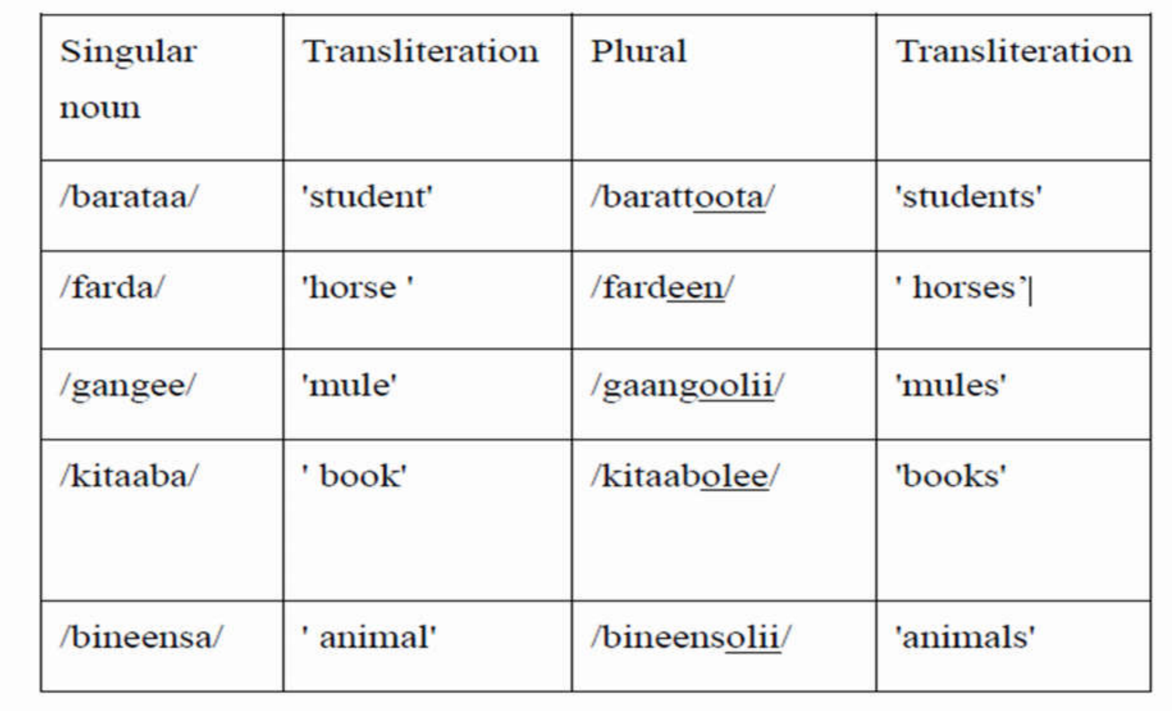
Adurreen maal **ariitii**? What does the cat run after?

II. Number

Afan Oromo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialect to dialect. In connection with numbers the plural suffix is very often considered unnecessary: **harka ishee lamaaniin** with her two hand(s).

The majority of plural nouns are formed by using the suffixes [56, 60]. – **oota**, followed by –**lee**, **-wwan** , **-een**, **-olii**/ **-olee** and –**aan**. Other suffixes like –**iin** in **sariin** (**dogs**) are found rarely.

Table 3.1: Afan Oromo Plural Noun Suffixes



Singular noun	Transliteration	Plural	Transliteration
/barataa/	'student'	/barattoota/	'students'
/farda/	'horse '	/fardeen/	' horses'|
/gangee/	'mule'	/gaangoolii/	'mules'
/kitaaba/	' book'	/kitaabolee/	'books'
/bineensa/	' animal'	/bineensolii/	'animals'

III. Definiteness

Definiteness is a grammatical category used for distinguishing noun phrases according to weather their reference in a given context is presumed to be uniquely identifiable. In Afan Oromo demonstrative pronouns like "**kun**" (this), **sun** (that) is used to express definiteness.

Mucaan kun this/ the Child (Subject)

Mucaa kana this/ the Child (Object)

Mucaan sun that/ the Child (Subject)

Mucaa sana that / the Child (Object)

To express indefiniteness emphatically the Oromo speaker my use numerical **tokko** one, Example: **Muka tokko** one / a Tree.

In some Afan Oromo dialects the suffix -**icha** (male), -**ittii**(n)(female) which usually has a singularize function is used where other languages would use a definite article.

Example:

Jaarsicha the old man (Subject) **Jaartittii** the old man (Subject)

IV. Derived Noun Forms

The most common word formation methods in Afan Oromo are derivational and compounding [56].

Derivation

Derivational suffixes are added to the root or stem of the word. From derived verbal stem and adjectives may be formed by means of derivational suffixes. The following suffixes play an important role in Afan Oromo word derivation. They are **-eenya, -ina, -ummaa, -annoo, -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata,** and **–ooma.**

Examples:

jabaa strong **jabeenya** strength

jabina strength **jabee** intensive

jabummaa strength **jabaachuu** to be strong

jabaachisuu to make strong **jabeessuu** to make strong

jajabaachuu to be consoled **jabeefachuu** to make strong for one self

V. Compound words

On the other hand, it seems that the use of genitive constructions is a very old method of forming compound nouns, as traditional titles shown.

abbaa gadaa traditional Oromo president

abbaa caffee chairman of the legislative assembly

abbaa dubbii chief speaker of the caffee assembly

abbaa duulaa traditional Oromo minister of war

3.2.1.2. Afan Oromo Verbs

Verbs are content words that denote an action, occurrence, or state of existence. Afan Oromo has base (stem) verbs and four derived verbs from the stem. Moreover, verbs in Afan Oromo are inflected for gender, person, number and tenses.

The four derived stems the formation of which is still productive in Afan Oromo are:

Autobenefactive (AS)

Passive (PS)

Causative (CS)

Intensive (IS)

Passive, causative, and autobenefactive are formed with addition of a suffix to the root, yielding the stem that the inflectional suffixes are added to. The personal terminations according to different conjunctions are added to these affixes. The intensive stem is formed by reduplicating the first consonant and vowel of the first syllable of the stem. The derived stems may be formed from all verbs the meaning of which permits it [56, 60].

1. **Autobenefactive**

The Afan Oromo autobenefactive (or "middle" or "reflexive-middle") is formed by adding **(a)adh**, **-(a)ach or -(a)at** or sometimes **-edh, -ech** or –**et** to the verb root.

This stem has the function to express an action done for the benefit of the agent himself.

Example: **bitachuu** to buy for oneself the root verb in this case is bit- The conjugation of a middle verb is irregular in the third person singular masculine of the present and past (-**dh** in the stem changes to -**t**) and in the singular imperative (the suffix is -**u** rather than–**i**).

Examples:

bit- buy **bitadh**- buy for oneself

Infinitive and participles are always formed with **-(a)ch,** while the imperative forms have **-(a)(a)dh** instead of **-(a)ch**.

Infinitive Imperative singular. Imperative plural. English

argachuu argadhu argadhaa to find/get

2. Passive

The Afan Oromo passive corresponds closely to the English passive in function. It is formed by adding -am to the verb root. The resulting stem is conjugated regularly.

Example: **beek**- know **beekam**- be known

3. Causative

The Afan Oromo causative of a verb corresponds to English expressions such as: cause, make, let. It is formed by adding **-s**, -**sis**, or -**siis** to the verb root example:

Deemuu to go **deemsisuu** to cause to go

4. Intensive

I. Simple tenses

It is formed by duplication of the initial consonant and the following vowel, geminating the consonant.
Example: **Waamuu** to call, invite **wawwaamuu** to call intensively

II. Infinite Forms

Infinite forms can be formed in two ways: Infinitive and Participle/gerund **Infinitive**

Infinitive is an uninflected form of the verb. In Afan Oromo infinitive form of verbs terminates in -**uu**. Examples:

arguu to see **deemuu** to go

On the other hand, the infinitive forms of autobenefactive verbs terminate in -**chuu**.

Example: **jiraachuu** to live **bitachuu** to buy for oneself

a) Participle/ gerund

Participle is a non-finite form of the verb whereas a gerund is a noun formed from a verb (in English the '**-ing'** form of a verb when used as a noun). In Afan Oromo a participle is formed by adding **-aa** to the verb stem [60].

Example:

deemaa going **jiraachaa** living

According to the meaning of the verb these forms may serve as agent nouns.

barsiisaa teacher **gaafatamaa** responsible person

For these agent nouns feminine forms are used according to the pattern of feminine adjective formation.

barsiiftuu teacher **gaafatamtuu** responsible person

On the other hand, a gerund is formed by adding **-naan** to the verb stem.

deemnaan after having gone **nyaannaan** after having eaten

b) Imperative

Imperative singular of base stems and all derived stems beside autobenefactive stems is formed by means of the suffix **-i**. Example:

deemi! go! **argi**! look!

The imperative singular of autobenefactive stems is formed by means of the suffix **-u**. Example:

jiraadhu! live!

Imperative plural of all stems is formed by means of -aa.

Example: **deemaa**! go! **argaa**! see!

Negative imperatives are formed by means of - (**i**)**in** for singular and -(**i**)**inaa** for plural.

Example: **Qubaan jechoota irra hin deemiin.** Don't point on the words with your finger.

c) Finite Forms

The Afan Oromo uses different conjugations for the verbs in main clauses and in subordinated clauses for actions in present or near future. The first-person singular is differentiated from the third person masculine by means of an -n that normally is suffixed to the word preceding the verb.

Present Tense Main Clause Conjugation

The present tense main clause conjugation is characterized by the vowel -a:

deemuu to go

1.p. S **deema**

2.p. **deemta**

3.p.m **deema**

3.p.f **deemti**

l.p. pl **deemna**

2.p. and polite form **deemtu/deemtan(i**)

3.p. and polite form **deemu/deeman(i**)

Examples: **gara mana barnootaan deema**. I go to the school.

Past tense conjugation

The past tense conjugation is characterized by the vowel -e:

deemuu to go

3.p.S **deeme**

33.p.m **deeme**

3.p.f **deemte**

l.p.pl **deemne**

2.p. and polite form **deemtani**

3.p. and polite form **deemani**

Example: **gammachiis gara mana yaalaa deeme.** gammachiis went to the school.

Subordinate Conjugation

The subordinate conjugation is used in affirmative subordinated clauses and in connection with the particle **akka** for the jussive. Beside this the subordinate conjugation is used to negate present tense actions.

Deemuu to go

l.p.S **akkan deemu**

2.p. **akka deemtu**

3.p.m. **akka deemu**

3.p.f. **akka deemtu**

l.p.pl **akka deemnu**

2.p. and polite form **akka deemtani**

3.p.and polite form **akka deemani**

Examples: **Akkan yaadutti barnooti jira**. As I thought there is a school.

Contemporary verb conjugation

The contemporary verb conjugation is used only in connection with the temporal conjunction **-odoo,-otoo,-osoo,-otuu** or **-utuu** that being connected with this conjugation means ‘while’. The contemporary verb conjugation is a kind of subordinated conjugation with lengthened final vowels.

Example: **"Otuun isin waamuu maaliif deemtu?" jedhe**. He said, "While I was calling you (pl.) Why do you go?"..p **deemte**

Jussive

To form the jussive in Afan Oromo the particle **haa** has to be used in connection with the subordinate conjugation.

Example: **Isaan haa deemani** they shall go

Negation iii. Verb Derivation

Present tense main clause actions are negated by means of the negative particle **hin** and the verb in subordinate conjugation.

Example: **Dastaan hin jiru**. Desta is not present.

Present tense actions in subordinated clauses are negated by means of the negative particle **hin** and a suffix -**ne** that is used for all persons. Past tense actions are negated in the same way using the particle **hin** and the suffix -**ne**.

Example: **Ani dhufuu hin danda'u**. I can't come

Some Afan Oromo verbs are derived from nouns or adjectives by means of an affix -**oom**. These verbs usually express the process of reaching the state or quality that is expressed by the corresponding noun or adjective. From these process verbs causative and autobenefactive stems may be formed.

Examples: **danuu** much, many, a lot **guraacha** black

danoomuu to become much **gurraachomuu** to become black

Causative verbs, however, can also be derived directly from adjectives or nouns by suffixing a causative affix -**eess** to the stem of the noun or adjective, example:

danuu much **daneessuu** to increase, multiply

Another means to derive process verbs from adjectives in Afan Oromo is to form an autobenefactive stem.

Example: **Adii** white **addaachuu** to become white

Compound Verbs

In addition to the above discussed derived verbs, compound verbs can be formed by means of pre-/postpositions, pronouns and adverbs in Afan Oromo such as ol above, gad below, wal, waliin, walitti, wajjin together, keessa in, jala under; they precede different verbs and express a broad variety of meanings [59].

Examples: gadi dhiisuu to let go of gaddhiisuu to let go of

Compound verbs can also be formed with jechuu or gochuu.

Example: With jechuu with gochuu

cal jechuu (to be quiet, silent) cal gochuu (to make quiet silent)

v. 'To be' and 'to have'

Afan Oromo has different means to express 'to be'. One of them is copulas, other means are the verbs ta'uu, jiruu and turuu.

The morphemes (-)dha and (-)ti (suffixed or used as independent words) serve as affirmative copulas as well as the vowel -i that is added to nouns terminating in a consonant. The copula dha is used only after nouns terminating in a long vowel.

Negative copula is miti, irrespective of the termination of the noun.

Examples:
Present tense: Anis jabaa dha. I am strong, too.
Nouns terminating in a short vowel do not take any copula.

Example: Isheen durba. She is a girl.
Nouns and pronouns terminating in a consonant are combined with the copula.

Example: Kuni bisbaani. This is water.
In all utterances related to possession only the copula -ti may be used.

Example: Hojiin hundee guddinaa ti! Work is the basis of development.
Present progressive:

Past Tense

Waa'een jarreen Axaballaa warra isaaniitiif qofa otuu hin taane uummata naannoofiyyuu hibboo ta'aa iira.

The life of Axaballaa is like a mystery not only, for his family, but also for the people around him.

Sangaan kan eenvuu ture? Whose ox was it?

The forms of the verb **qabuu** 'to have' are overlapping with the forms of the verb **qabuu** 'to grasp', 'keep'.

The verb **qabuu** appears with the meaning 'to have' only in the present tense and one past tense form. In present tense conjugation both verbs have the same form.

3.2.1.3. Afan Oromo Adjectives

An adjective is a word which describes or modifies a noun or pronoun. A modifier is a word that limits, changes, or alters the meaning of another word. Unlike English adjectives are usually placed after the noun in Afan Oromo. For instance, in **Tolaan farda adii bite** "Tola bought white horse" the adjective **adii** comes after the noun **farda**. In Afan Oromo sometimes it is difficult to differentiate adjective from noun [58].

Example:

dhugaa truth, reality, true, right

dhugaa keeti your truth/ you are right (truth served as noun)

obboleessi hiriyaa dhugaati brother is the friend for truth / brother is a true friend (true served as adjective)

I. Gender

In Afan Oromo adjectives are inflected for gender. We can divide adjectives into four groups with respect to gender marking. These are:

a. In the first group the masculine form terminates in –**aa**, and the feminine form in –**oo**.

Example:

guddaa (m.) **nama guddaa** a big man

guddoo(f.) **nama guddoo** a big woman

b. In the second group the masculine form terminates in –**aa**, the feminine form in –**tuu** (with different assimilations).

Example: **dheeraa**(m.) **nama dheeraa** a tall man

dheertuu(f.) **intala dheertuu** a tall girl

c. Adjectives that terminate in –eessa or –(a)acha have a feminine form in –eettii or –aattii.

Example: **dureessa** (m.) **nama dureessa** a rich man

dureettii (f.) **nitii dureettii** a rich woman

Adjectives whose masculine form terminates in a long vowel other than –aa as in short vowel –a (but not of the suffix –eessa/-aacha) are not differentiated with respect to their gender.

collee(m.) **farda collee** an active horse

collee(f.) **gaangee collee** an active mule

II. Number

There are four groups of adjectives with respect to number. These are:

a. Most of the adjectives form the plural by reduplication of the first syllable masculine and feminine adjectives differ in plural as they do in singular [59].

Example:

Singular Plural

guddaa(m.) **guguddaa**(m.)

guddoo(f.) **guguddoo**(f.)

xinnaa(m.) **xixinnaa**(m.)

xinnoo xixinnoo

pl.f. **lageewwan guguddoo** big rivers

pl.m. **qubeewwan guguddaa fi xixiqqaa** big and small letters

b. There is a further plural form which is gender neutral for adjectives of this group beside a special masculine and feminine plural. This plural form terminates in **-oo**, and is sometimes used with reduplication and sometimes without. Table 4 shows examples of plural adjectives formed by reduplication which are gender neutral

Singular plural plural

Jabaa (M) **Jajabaa**(M) **Jajjaboo**(Gender neutral)

Jabduu (F) **Jajjabduu**(F)

c. Adjectives which may function as nouns as well form the plural only by using noun plural suffixes. Table 5 shows examples of plural adjectives formed using noun plural suffixes

Singular Plural

M F M F

Dureessa Dureettii Dureeyyii/dureessota dureettiwwan

d. Adjectives of the fourth group form the plural without marking the gender, very often by reduplication of the first syllable. Sometimes adjectives of this group form the plural by using a noun plural suffix [56].

Singular Plural English

Adii a`adii/adaadii White

Collee Colleewwan Active

III. Definiteness

The demonstrative pronouns that express definiteness in Afan Oromo follow the adjective if the noun is qualified by an adjective and a demonstrative pronoun as well.

Example: **Namicha dheeraa sana argitee**? Did you see that tall man?

The suffix –**icha** that sometimes has a definite function normally is suffixed to nouns, but it can be suffixed to adjectives or numerals, too,

Example **Lagni guddichi** the big river **namichi tokkichi** a single man

IV. Compound Adjectives

In the new terminology of Afan Oromo compound adjectives play a growing role.

Example: **afrogaawaa afur + rogaawaa** rectangular four + angled

sibilala sibila + ala non-metal metal + outside

3.2.1.4. Adverbs

Adverbs have the function to express different adverbial relations such as relations of time, place, and manner or measure.

Some examples of adverbs of time:

amma now

booda later

Some examples of adverbs of place:

achi(tti) there

ala outside

Some examples of adverbs of manner:

saffisaan quickly

sirritti correctly

Some examples of adverbs of measure:

baay'ee , **danuu** much , many , very

duwwaa only, empty

3.2.1.5. Pre-, Post, and Para-positions

Afan Oromo uses prepositions, postpositions and para-positions [58].

I. Postpositions

Postpositions can be grouped into suffixed and independent words.

Suffixed postpositions

-tti in, at, to

-rra/irra on

-rraa/irraa out of, from

The post position –**tti** is used to form the locative. The postposition **-rraa/irra** may be used to express a meaning similar to ablative.

Example: **Adaamaatti yoom deebina?** When shall we go back to Adama?

Gammachuun sireerra ciise. Gemachu lay down on bed.

Post position as independent words

ala outside **wajjiin** with, together with

bira beside **teellaa** behind

Example: Namoota nu bira jiraniis hin jeeqnu. We don't hurt people who are with us.

II. Prepositions

akka like, according to

gara to, in the direction of

hanga/hamma until, up to

karaa along, the way of, through

The prepositions **gara**, **hanga**, and **waa'ee/waayee** are still treated as nouns and therefore are used in a genitive construction with other noun they belong to, expression: the direction to, the matter of, etc.

Example:

Namni akka harkaan waa hojjechuuf fayyadamu argi maalitti fayyadamaa? As people use hands to work something what does the elephant use?

III. Para-positions

Gara… tti to **Gara… tiin** from the direction of

Example: **Lukkichi rifatee jeedaloo dheesuuf gara manaatti gale.** The cock was scared and went home to take refuge from the fox.

3.2.1.6. Conjunctions

Conjunctions are unchanging words which coordinate sentences or single ports of a sentence. The main task of conjunctions is to be a syntactical formative element that establishes grammatical and logical relation between the coordinated constituents.

According to [46] the main functions of conjunctions are identified as: the function of coordinating clauses (coordination), the function of coordinating parts of sentence (coordination) and the function of coordinating syntactical unequal clauses (subordination). On the other hand, with regard to their form we can subdivide the conjunctions of Afan Oromo into:

Independent Conjunctions

a) **Coordinating**

Example: **garuu** but

Hoolaan garuu rooba hin sodaattu. But the sheep is not afraid of rain.

b) **subordinating**

Example: **akka** that, as if, as whether

Maaliif akka yaada dhuunfaa yookaan yaada haqaa akka ta'e adda baasii barreessi. Write separately why it is an individual opinion or that it is an opinion about justice

I. Suffixed Conjunctions

Example: **–f/ -fi/ -dhaaf** and, that, in order to, because, for

uffata uffachuuf bittee? Did you buy the clothe for wearing?

II. Conjunction consisting of one, two or more parts

Conjunctions consisting of two parts can be formed by two independent words or two enclitics or one independent word plus enclitic. They can be formed made up of two single conjunctions that are used after each other in order to give more detailed information about the logical relation or to intensify it.

Example: **akkam akka** how, that

Dura namni tokko beekumsa mammaaksaa akkam akka jabeeffatu ilaaluu nu barbaachisa. At first, we have to see how a person extends the knowledge of proverbs

III. Conjunctions consisting of several segments

Conjunctions consisting of several segments are copulative or disjunctive conjunctions which as they stand separately from each other are to emphasize the segments of a parallel construction. These are stable, stereotyped constructions the first segment of which has to be followed by a certain second segment:

Example: **–s… -s,** as well as

Jechoota hudhaa wajjiiniis, hudhaa malees karaa lamaan barreeffaman

Words with glottal stop as well as without glottal stop are written in two ways.

3.2.2. Word and Sentence Boundaries

In Afan Oromo, like in English the blank space shows the end of a word. Moreover, parenthesis, brackets, quotes, etc. are being used to show a word boundary. Sentence boundaries punctuations are also similar to English language i.e., a sentence may end with a period (.), a question mark (?), or an exclamation point (!) [59].

Morphology adds a burden to NLP works. For the purpose of text summarization and also other NLPs, the variant words of a morpheme should be reduced to their root so that they can be counted as one while calculating term frequency, and in our case when creating a Word2Vec model. Using stemmer is believed to minimize the difficulty of dealing with different forms of a word [59]. There have been efforts of developing stemming algorithm for Afan Oromo. We used the algorithm developed by [59]. For our work.

CHAPTER FOUR

4. METHODOLOGY

4.1. Introduction

A research methodology is a way to systematically solve the research problem. In this section, the procedure for undertaking an extractive and an abstractive text summarization for Afan Oromo Proclamation texts is presented. It included the corpus preparation, data description, methods, procedures, the model and the evaluation techniques which are presented respectively as follows.

4.2. The Methods

The method followed is the summarization of the extractive and the abstractive Afan Oromo proclamation summarizations. In this study the Recurrent Neural Network (RNN) Algorithm with Long Short Term Memory (LSTM) for abstractive text summarization and TextRank algorithm for extractive text summarizations were used.

4.3. Data Description

The dataset of 583 articles of 27 Afan Oromo Proclamations were used for the experimentation purpose.

4.3.1. Corpus Preparation

Afan Oromo proclamation articles were taken for the dataset. The data are used to develop the pretrained model and for experimentation (validation and testing). The data needed was collected to develop the model. The digitally posted data is collected from web pages of oag.gov.et and/or www.caffeeoromiyaa.org in scanned pdf format. The proclamations are presented in three languages, namely, Afan Oromo, Amharic, and English on the "Magalet of Oromia" Magazine which is one of the popular magazines in Oromia Regional State. We have converted the scanned pdf format into word document format by OCR application software. Twenty seven Afan Oromo Proclamations and their corresponding 583 articles that were used for the dataset are displayed in table 4.3.1.

Table 4.3.1: Proclamations and articles used as dataset

Proclamations	*Articles*
Labsii Lak 18 Bara 1989 / Proclamation Nº. 18/1997	10
Labsii Lak 24 Bara 1990/ Proclamation Nº. 24/1998	4
Labsii Lak 28 Bara 1991/ Proclamation Nº. 28/1999	11
Labsii Lak 35 Bara 1992/ Proclamation Nº. 35/2000	7
Labsii Lak 48 Bara 1994/ Proclamation Nº. 48/2001	9
Labsii Lak 72 Bara 1995/ Proclamation Nº. 72/2003	18
Labsii Lak 112 Bara 1998/ Proclamation Nº. 112/2006	38
Labsii Lak 141 Bara 2000/ Proclamation Nº. 141/2008	34
Labsii Lak 163 Bara 2003/ Proclamation Nº. 163/2011	33
Labsii Lak 179 Bara 2005/ Proclamation Nº. 18/2013	19
Labsii Lak 224 Bara 2012/ Proclamation Nº. 18/2020	20
Labsii Lak 79 Bara 1996/ Proclamation Nº. 18/2004	92
Labsii Lak 112 Bara 1995/ Proclamation Nº. 112/2003	73
Labsii Lak 50 Bara 1994/ Proclamation Nº. 50/2002	61

4.4. Data Preprocessing

Data preprocessing step is one of the basic important tasks before proceeding with the model-building part. To avoid the catastrophic move towards data building part, having clean data that is free of distortion and ambiguities is very important. Therefore, all the unwanted symbols, characters, etc. that do not affect the objective of the summarization are dropped. The following important preprocessing tasks were performed for the desired results.

- All the texts were converted to lower case;
- Every text inside the parenthesis ( ) was removed;
- Punctuations and special characters were eliminated;
- Stop words were removed; and

- Short words or abbreviations were also removed.

In cleaning the data process apostrophe plays an important role in reading and writing system the Afan Oromo language. There are times when the apostrophe is used interchangeable with the spelling “h”. For instance, “ka’e”, “ba’aa”

4.5. Procedures

The analysis of the data followed the appropriate models and techniques for each method in deep learning approach for both the extractive and abstractive text summarization methods.

4.5.1. Abstractive text summarization

4.5.1.1. Recurrent Neural Network model (RNN) Algorithm

The RNN algorithm along LSTM model networks which are the better versions of RNN are used [61]. One of their benefits is they resolving the vanishing gradient problem to remember the past data. To train the model LSTM uses back propagation in text summarization methods. They [61] continued to state that for abstractive summarization it is common to use either GRU or LSTM cells for the RNN encoder and decoder. The use of LSTM cells was preferred for its extra control via their memory unit, although many top models use GRU cells for their cheaper computation time [61]. Hence, in this study, LSTM algorithm was used for the Afan Oromo proclamation abstractive text summarization.

4.5.1.2. Sequence-to-Sequence Model

The sequence-to-sequence model is used to convert the sequences of one input domain into the sequences of another output domain in the Afan Oromo proclamation texts. The Sequence to Sequence model uses a method of encoder-decoder architecture, which is a method of encoder-decoder-based machine translation. Its function is to map an input of sequence to an output of sequence with a tag and attention value. Here the two LSTMs that work together with a special token are used to predict the succeeding state sequence from the previous sequence.

An encoder and a decoder which are the parts of a standard sequence-to-sequence model are used in this study to be forwarded to a network of decoders to produce the sequence representing the output. As the attention mechanism that highlights relevant features from the Afan Oromo proclamation text is also used. The main characteristic behind attention is to measure a weight distribution on the input data, assigning more important elements to higher values. [61]. whereas context vector defines as the sum of hidden states of the input sequence for weighted alignment scctore.

4.5.1.3. Encoder-Decoder architecture

In the case of varying input and output data length, encoder-decoder architecture is used to predict the sequence. The encoder reads the entire input sequence and a fixed-length internal representation is generated. The entire context of the input data sequence is captured by the internal representation. Then model, that is, sequence-to-sequence, is applied to the tokens [54].

The encoder consisted has three stacked layers of LSTM whereas the decoder has one LSTM layer. Next, the attention layer receives the encoder and decoder's output [61]. The attention layer's output is then concatenated with the decoder output and fed to a time-distributed dense layer. The decoder network uses this internal representation to predict the output words until the end of the sequence token is reached.

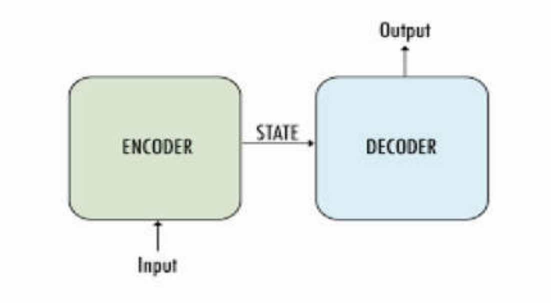


Figure 4.1.: *Encoder-Decoder*
Source: [61] Nallapati, et. al.

Encoder

An encoder is an LSTM network which reads the entire input sequence that is the Afan Oromo proclamation texts dataset. At each time step, one word from the input sequence is read by the encoder [54, 61]. It then processes the input at each time step and captures the context and the key information related to the input sequence. It takes each word of input(x) and generates the hidden state output (h) and the cell state which is an internal state(c). The hidden state (hi) and cell state (ci) of the last time step are the internal representation of the complete input sequence which was used to initialize the decoder.

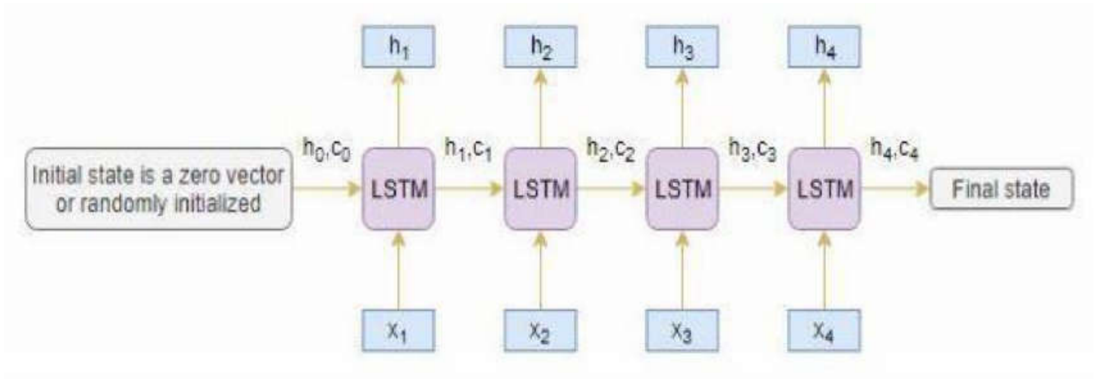


Figure 4.2.: Encoder
Source: [61] Nallapati, et. al.

Decoder

The decoder is also an LSTM network. It reads the entire internal representation generated by the encoder one word at a time step. It then predicts the same sequence offset by a one-time step. The decoder is trained to predict the next word in the output sequence given the previous word based on the contextual memory stored by the LSTM architecture [54]. Before feeding the token to the decoder, the two special tokens <start> and <end> are added at the beginning and the end of the target sequence. Then the target sequence is predicted by passing one word at a time. The first word of output of the decoder is always <start> token. The end of the output sequence is represented by <end> token.

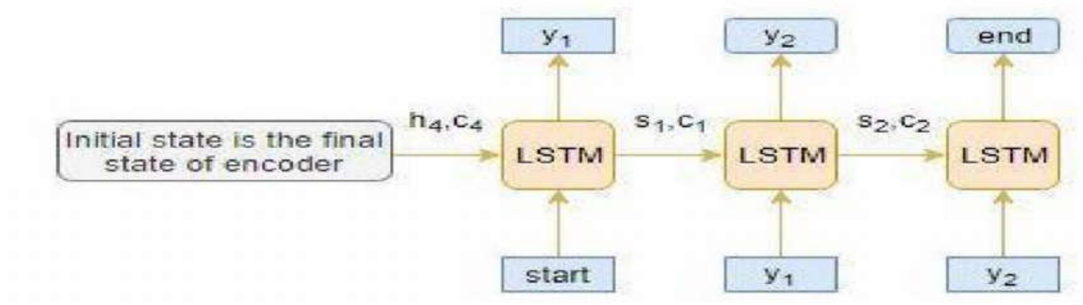


Figure 4.3.: *Decoder*
Source: [61] Nallapati, et. al.

The above architecture of the model is built using the TensorFlow library which is used to build layers in neural networks. The final architecture of the model will be as shown below.

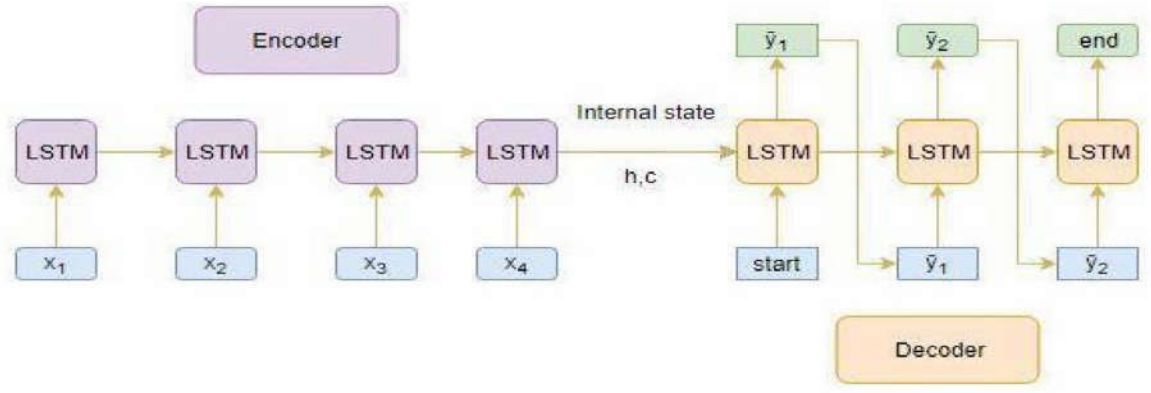


Figure 4.4.: LSTM Seq2Seq model architecture
Source: [61] Nallapati, et. al.

4.5.1.4. Attention Layer

A Sequence to sequence model with an attention mechanism consists of an encoder, decoder, and an attention layer [44]. Sep, et.al. [44] continues to explain that the attention mechanism is used to secure individual parts of the input which are more important at that particular time. It can be implemented by taking inputs from each time step and giving weightage to time steps. The weightage depends on the contextual importance of that particular time step. It helps pay attention to the most relevant parts of the input data sequence so that the decoder can optimally generate the next word in the output sequence. This study employed a sequence to sequence model with the attention mechanism that contain encoder decoder and the attention layer.

For encoder-decoder neural networks, the use of attention allows for the creation of a context vector at each timestep, given the decoder's current hidden state and a subset of the encoder's hidden states. For global attention, the context vector is conditioned on all of the encoder's hidden states, whereas local attention uses a strict subset of the encoder's hidden states. There are two different scoring functions, which are used as the weights when averaging the hidden states to produce the context vector.

Score $(h_t, h_s) = h_t{}^{T}\ h_s$ (dot product form) score $(h_t, h_s) = h_t{}^{T}\ W_a\ h_s$ (bilinear form). The scoring functions are then used as follows:

$$a_t(s) = \mathrm{softmax}(s)$$

Applying the softmax function to the raw scores creates a probability distribution over the encoder's hidden states, which are then used to create the context vector as follows:

$$c_t = \sum_{!} a(s)h$$

c_t is a context vector at time step t with the same dimensionality as the decoder hidden states. Finally, we use h_t and c_t to compute the next hidden state in the decoder, h_{t+1}:

$$h_{t+1} = \tanh(W[h_t\ ;\ c_t] + b)$$

4.5.1.5. Tokenization

Tokenization is a form of fine grained data protection that replaces a clear value with randomly generated synthetic value which stands for the original as a 'token' [61]. The pattern for the tokenized value is configurable and can retain the same format as the original which means les down-stream applications changes, enhanced data sharing, and more meaningful testing and development with the protected data. The keys are used in the tokenization of words so that the input to the model is a set of numbers rather than words so as to make the computation easier using vectors [61]. Tokenizing data is important as networks need numerical data to work on rather than raw data with characters.

4.5.1.6. Handling Varying Sequence Lengths

Because input and output sequence lengths vary in the current dataset, the researcher standardized all inputs to be some max length by padding short sentences with PAD tokens and cutting off longer sentences. This padding was taken into account in both the scoring and loss functions, preventing the model from updating parameters that should not be updated.

4.5.1.7. Generating Summaries

At test time, the model's decoder feeds its output as a word embedding input into the next decoder cell. For an LSTM, the output at time $t = h_t$. Because using hidden states with dimensionalities equal to the size of the model's vocabulary would take too long to train. The outputs of the LSTM decoder cells are actually dense vector representations of the final outputs [65]. By using an affine transformation in the decoder from output h_t, the prediction of the word generated predictions as follows:

$$y_t = Wh_t + b \ W \in R(V, H)$$

Projecting the dense vector through the affine transformation allows us to create a probability across the entire vocabulary size, and then take an argmax in order to be able to index into the embedding matrix and retrieve the input for the next time step.

4.5.2. Extractive text summarization

4.5.2.1. Algorithm

TextRank is an extractive text summarization technique. It takes as input the summaries and splits the whole summary into individual sentences [65]. TextRank algorithm is used to identify the most important sentences in a text based on information exclusively drawn from the text itself. Unlike other supervised systems, which attempt to learn what makes a good summary by training on collections of summaries built for other articles, TextRank is fully unsupervised, and relies only on the given text to derive an extractive summary, which represents a summarization model closer to what humans are doing when producing an abstract for a given document.

R. Mihalcea, et. al. [64] discussed the important aspect of TextRank that it gives a ranking over all sentences in a text – which means that it can be easily adapted to extracting very short summaries. They [64] also added that the other advantage of TextRank over previously proposed methods for building extractive summaries is the fact that it does not require training corpora, which makes it easily adaptable to other languages or domains. TextRank algorithm is used for the process of extractive Afan Oromo text summarization. In the process of identifying important sentences in a text of the Afan Oromo proclamation dataset, a sentence recommends another sentence that addresses similar concepts as being useful for the overall understanding of the text. The sentences that are highly recommended by other sentences in the text are likely to be more informative for the given text, and will be therefore given a higher score.

Lastly, it calculates the similarities between each sentence embedding and stores it in a matrix. For sentence rank calculation, the similarity matrix is then transformed into a graph, where the sentences are vertices, and similarity scores are edges.

4.5.2.2. Cosine similarity

Cosine similarity is used to re-arrange sentence extraction from the result of keyword extraction algorithm process [65]. The keyword extraction algorithm using calculation based on TF/IDF, weight a given term to determine how well the term describes an individual document within a corpus [65]. It does this by weighting the term positively for the number of times the term occurs within the specific document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term t and document d, where t appears in n of N documents in D. The TF-IDF function is of the form as follows:

$$TFIDF\ (t,dn,,N)= TF\ (t,d)\ x\ IDF\ (n,N)$$

When the TF-IDF function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher TF-IDF score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is significant to the document, and could be used to accurately summarize the document. TF-IDF provides a good heuristic for determining likely candidate keywords, and it (as well as various modifications of it) has been shown to be effective after several decades of research.

Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining [65]. Given two vectors of attributes, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as:

$$Similarity = Cos\ \theta = \frac{A.B}{|A||B|}$$

The resulting similarity ranges from 0 with usually indicating independence, and 1 with usually indicating exactly the same and in between those values indicating intermediate similarity and dissimilarity. For the text matching, the attribute vector A and B are usually the term frequency vectors of the documents.

4.7. The Evaluation Metrics

ROUGE- N is the evaluation metrics used to validate the comparison of the extractive and abstractive summaries [52]. ROUGE-N stands for ROUGE with N-gram Co-Occurrence Statistics. It measures an n-gram recall between reference summaries and their corresponding candidate summaries. Formally, ROUGE-N can be calculated as:

Formally, ROUGE-N can be calculated as:

$$ROUGE - N = \frac{\sum_{S \in \{Ref\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Ref\}} \sum_{gram_n \in S} Count(gram_n)},$$

where $Re\ f$ is the reference summaries and n represents n-gram. $Countmatch$ ($gramn$) represents the maximum number of n-grams in the reference summaries and the corresponding candidates. The numerator of ROUGE-N is the number of n-grams owned by both the reference summaries and the automatically generated ones, while the denominator is the total number of n-grams occurring in the golden summary. The denominator could be set to the number of candidate summary n-grams as well to measure the precision. However, ROUGE-N mainly focuses on quantifying recall, so precision is not calculated here.

ROUGE-1 and ROUGE-2 are the special cases of ROUGE-N that are usually chosen for the validation of the comparison of the extractive and abstractive summary outputs of the Afan Oromo Proclamations articles, The number 1 and 2 represent unigram and bigram, respectively. ROUGE-1, ROUGE-2 or ROUGE-N are adopted by most of the research works [52].

The uses of recall and precision in the context of ROUGE are explained by Lin [66]. Recall indicates how much of the reference summary (the original dataset) the system summary (machine generated summary) is recovering or capturing. It is computed as:

$$\frac{\text{Number of overlapping words}}{\text{Total words in the reference summary}}$$

Much of the words in the system summary may be useless of overloaded with unnecessary words. To solve this problem precision is used [66]. In terms of precision, what you are essentially measuring is how much of the system summary was in fact relevant or needed? Precision is measured as:

$$\frac{\text{Number of overlapping words}}{\text{Total words in the system summary}}$$

CHAPTER FIVE

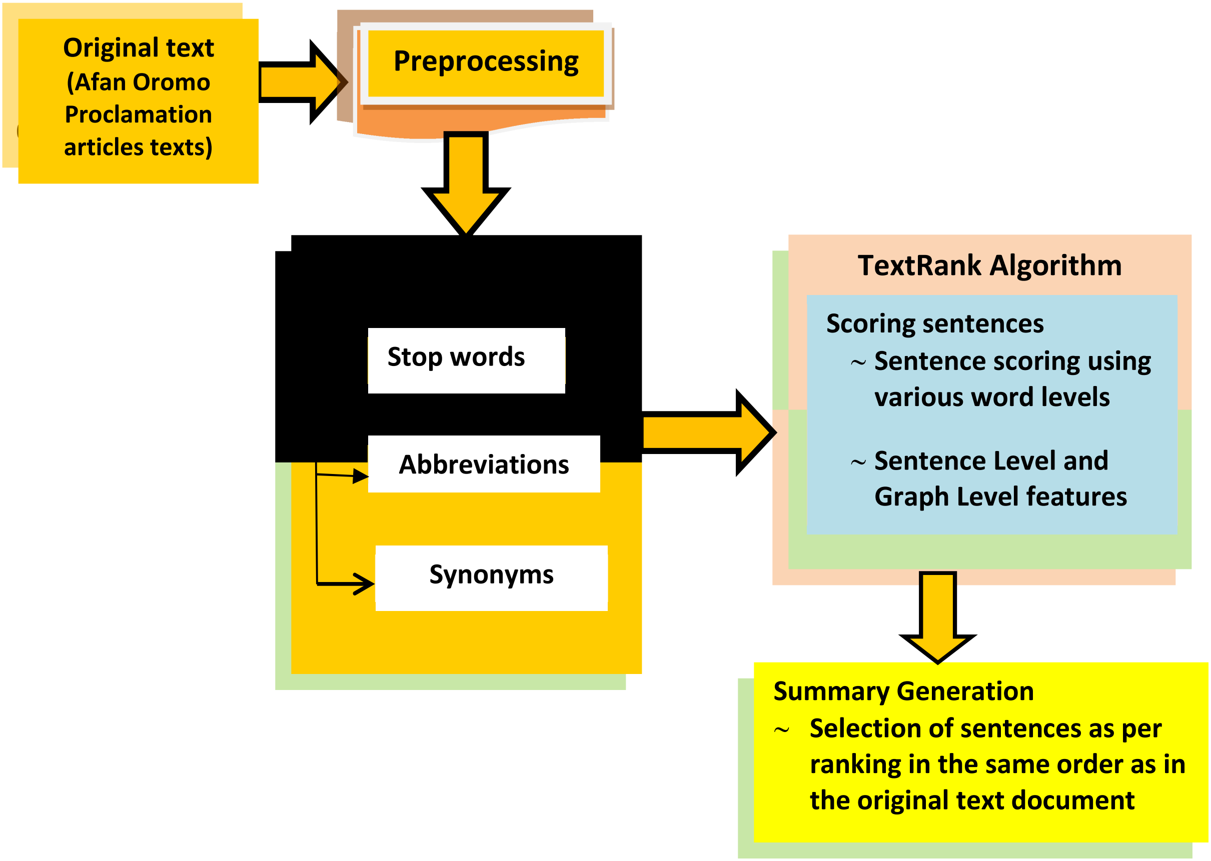
5. DESIGN, IMPLEMENTATION, EXPERIMENTAL RESULTS AND DISCUSSION

In this section design of the system, implementation, results of the experiment and discussions are portrayed.

5.1. System Design and Architecture

The system design and architecture is framed for the implementation of both the extractive and abstractive Summarizations and are presented as follows.

5.1.1. System Design and Architecture for Extractive Text Summarization



Original text (Afan Oromo Proclamation articles texts)
Preprocessing
Stop words
Abbreviations
Synonyms
TextRank Algorithm
Scoring sentences
~ Sentence scoring using various word levels
~ Sentence Level and Graph Level features
Summary Generation
~ Selection of sentences as per ranking in the same order as in the original text document

Figure 5.1.1: System design and architecture for extractive summarization

5.1.2. System Design and Architecture for Abstractive Text Summarization

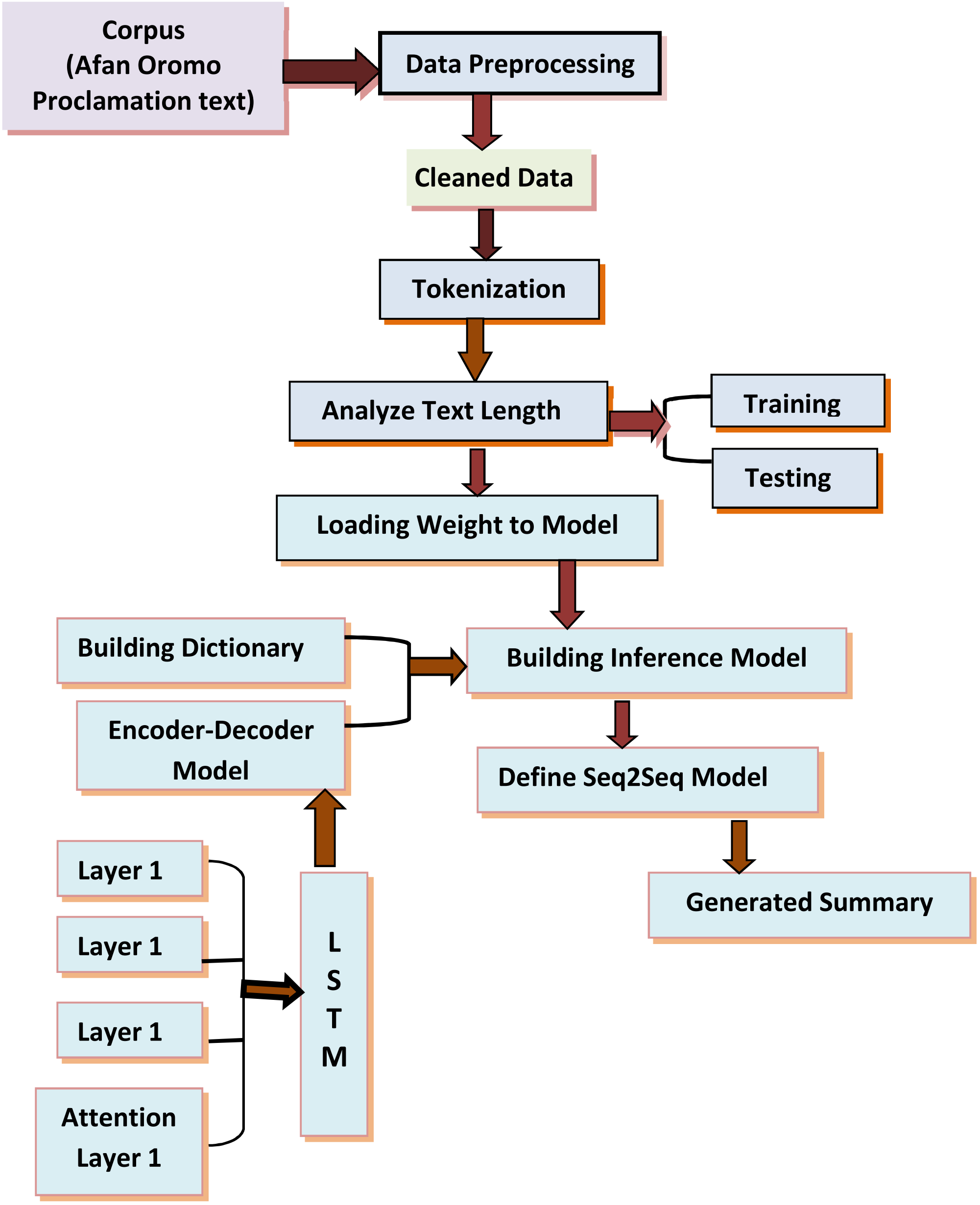


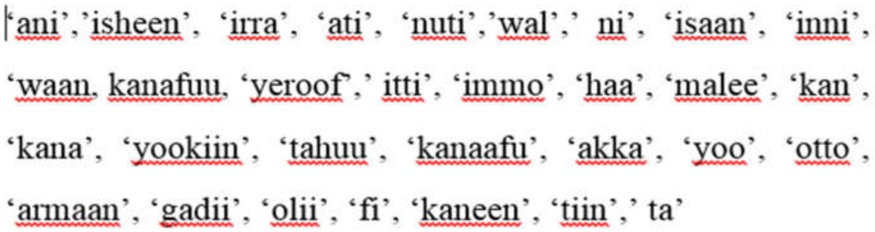
Figure 5.1.2: System design and architecture for abstractive summarization

5.2. Implementation

The experiment implementation procedures for the texts summarization were undertaken by cleaning the data and training the model which is followed by text summarization process.

5.2.1. Preprocessing

In preprocessing step was started by cleaning the data that consisted of the following steps: namely, stopword removal, lower the string, removing text inside parenthesis, removing numbers and other punctuations and special characters to generate cleaned summary of the dataset. In stopword removal frequently occurring uninformative words are removed because it doesn’t affect rather it speed up the processing. Figure 5.2 shows some of the Afan Oromo stopwords.



‘ani’,’isheen’, ‘irra’, ‘ati’, ‘nuti’,’wal’,’ ni’, ‘isaan’, ‘inni’,
‘waan, kanafuu, ‘yeroof’,’ itti’, ‘immo’, ‘haa’, ‘malee’, ‘kan’,
‘kana’, ‘yookiin’, ‘tahuu’, ‘kanaafu’, ‘akka’, ‘yoo’, ‘otto’,
‘armaan’, ‘gadii’, ‘olii’, ‘fi’, ‘kaneen’, ‘tiin’,’ ta’

Figure 5.2: Some Afan Oromo stopwords

One of the biggest challenges in analyzing textual data is how to change the unstructured data to a structured format in the preprocessing phase.

5.2.2. Training the model

The model is trained on the dataset (Afan Oromo articles texts). Among available dataset 80% is used for training and 20% is for validation purpose with 50 numbers of iteration (epochs). Figure 5.2 indicates the training loss and validation loss about how well the model is learning or how well the model is generalizing. In the process of training the data three outputs could come to the scene: under-fitting, fitting well or over-fitting of the model. There are training loss and validation loss. The training loss shows the extent to which the model fits the training data or the extent to which the model is learning Afaq, et.al. [67] and they also continued to explain that validation loss depicts that the extent to which the model fits the new data or how well the model is generalizing. A good fit of a model to summarized dataset is realized when the decrease to the point of stability occurs with the minimal gap between the final training and validation loss values. Hence, the graph in figure 5.2 indicates that there is a good fit due to a decrease to the point of stability with minimal gap between the two final values.

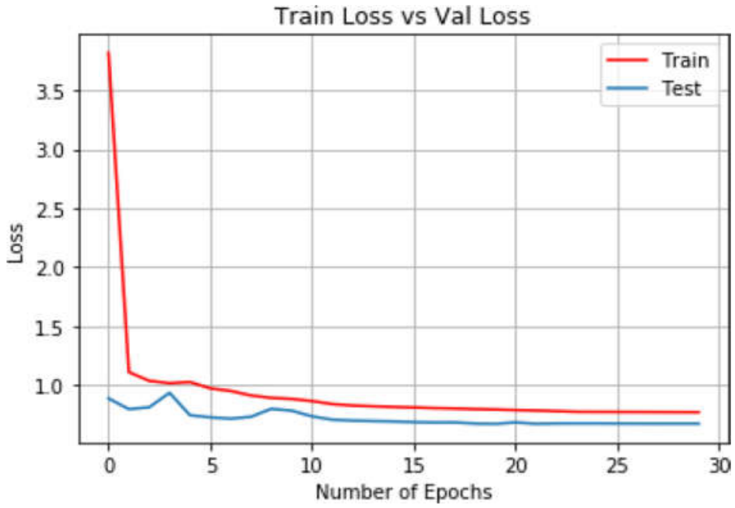


Figure 5.2.1: Graphical representation of train loss versus validation loss

5.2.3. Afan Oromo proclamation Abstractive text summarization

After the data is cleaned, the Afan Oromo proclamation abstractive text summary is generated by the model as follows.

(1) Adds a START and END token to the summary

(2) Analyzes the length of text and populate the lists with sentence lengths

(3) Use sklearn to split reviews data into training and testing

(4) Convert text sequences into integer sequences by preparing tokenization

(5) Build the Encoder-Decoder Neural Network architecture using LSTM layers

(6) Load the weights of the model trained

(7) Defining a function that uses the above encoder and decoder model to generate summary

(8) Generate summary

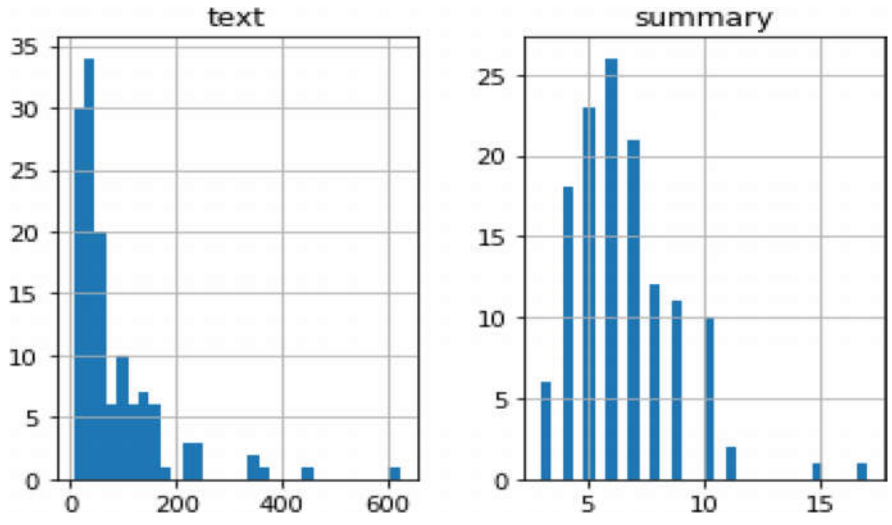


Figure 6.2.1: Graphical representation of maximum number of text and summary generated

5.2.4. Afan Oromo Proclamation Extractive Text Summarization

The summarization of the Afan Oromo proclamation texts extractive summarization was undertaken by the use of TextRank algorithm. The reason for selecting the algorithm is that for unsupervised approach for keywords and sentence extraction TextRank had achieved accuracy and its application is competent [64]. Hence, the summarization process was undertaken as follows.

The model:

1. Reads text and tokenize
2. Create vectors and calculate cosine similarity between two sentences
3. Sorts the rank and place top sentences
4. Output the summarized text or generate summary

5.3. Experimental Results

Extractive and abstractive Afan Oromo proclamation text summarization methods were compared in this study. Textrank algorithm was used for extractive text summarization method while Long Short Term Memory (LSTM) algorithm model which is a type of Recurrent Neural Network model was used for abstractive text summarization.

Sample Summaries for the abstractive Afan Oromo proclamation texts

The abstractive summarization of the texts has three steps, namely, review, original summary and predicted summary.

Sample one:

Review: seeraa hojii abbaa bilisummaa raawwatan seeraan haala biraatiin abba an seeraa kamiyyuu haallan kanaa gaditti ala umriin seeraan murtaae osoo dura hojii abbaa seerummaa hin gumiin bulchiinsa abbootii seeraa seera abbootii se eraatti badii yokiin hanqina dandeettii hojii cimaa qaba cimaa qaba jedhee yo kiin abbaan seeraa sababa fayyaa hojiisaa qajeellootti raawwachuu jedhee yero o murtiin gumichaa caffee naannichaatiif dhiyaatee sagalee caalmaa keewwata k eewwata xiqqaa tti gumichi abbaan seeraa tokko hojii abbaa isa jedhee amane d himmichi caffeedhaan murtaautti hojii abbaa yeroon abbaa seeraa kamiyyuu

Original summary: waaee abbaa aadaa fi

Predicted summary: mirga

Sample two:

Review: manni maree gandaa aangoofi hojjiiwwan gaditti ilaalaman qabaata kar ooraawwaniifi qajeelfamoota manni maree aanichaafi bulchiinsa aanaa hojiirra hordofa qajeelfamoota baasa godha maree gandichaa durataaa bulchaa gandichaa mana maree bulchiinsa durataaa bulchaa gandaatiin muudama aanaa bulchaa gandi chaafi abbootii seeraa mana murtii hawaasummaa raggaasisa karooraawwaniifi sa gantaalee hawaasummaa misooma diinagdeefi bulchiinsaa gochuudhaan sagantaa ra awwannaa hojii baasa hojiirra ooluu isaaniis hordofa ummata gandichaatiif bir o baasa hojiirra ooluu isaaniis hordofa ummata gandichaa hojii misoomaatiif k akaasa hojii kunuunsa qabeenya uumamaa hordofa gandicha keessatti sirni kabaj amuusaa mirkaneessa

Original summary: aangoofi hojii mana maree gandaa

Predicted summary: mirga hojii

Sample three:

Review: manni maree bulchiinsa naannichaa akkaataa aangoo heera mootummaa ke ewwata keewwata xiqqaa irratti tumaa yeroo ariifachiisaa baasa manni maree bu lchiinsa naannoo tumaa yeroo ariifachiisaa baasee hojiirra erga booda guyyaa kudha shan keessatti caffee walgaii qaba tumaan yeroo ariifachiisaa manni mar ee bulchiinsaa baase caffee fudhatama argate turu dandau jia qofaaf taa caffe en sagalee harka sadii keessaa yeroo murteessu tumaan yeroo ariifachiisaa jia afur dandaa tumaatni yeroo ariifachiisaa manni maree bulchiinsa mootummaa naa nnichaa caffeen tarkaanfiiwwan gama iyyuu mirgoota heera mootummaa keewwata irratti tauu

Original summary: tumaa yeroo ariifachisaa

Predicted summary: mirga

Sample four

Review: hojii olaanaa bulchiinsichaati tumaan keewwata keewwata xiqqaa irrat ti akkuma eegametti taee dura taaa bulchaan godinaa aangoo hojiiwwan kanaa ga ditti tarreeffaman niqabaata pireezidaanticha bakka buuudhaan godinicha nibul cha olaantummaan nihoggana sochii jaarmayoota adda addaa godinicha keessatti argamanii hordofa gabaasa addaa nikenna humnoota poolisii nageenya naannoo ee guufi seeraafi sirna godinichaa kabachiisuuf hundeeffaman olaantummaan nihogg ana nitoata hojii sadarkaalee bulchiinsaa biro godinicha keessatti qindeessa sochii hojii godinichaa pireezidaantichaafi mana maree bulchiinsa naannichaat iif yeroo yerootti gabaasa dhiheessa hojiiwwan biro pireezidaantichaafi mana maree bulchiinsa naannichaatiin kennamaniif raawwata

Original summary: aangoofi hojii duratаaa bulchaa godinaa

Predicted summary: mirga hojii

A. Sample summaries for the extractive Afan Oromo proclamation texts

The extractive summarization of the texts has two steps, namely, original and summary.

Sample 1

Original: Nannoon Oromiyaa

Qubsuma lafa walqabate kan ummatni Oromootii fi ummatootni Oromiyaa keessa jiraachuu filatan kan birootiis irra qubatan ta'ee;

Kaabaan- naanoo Affaarii fi naannoo Amaaraatiin,

Kiibbaan- naannoo Ummatoota kiibba Itoophiyatii fi keeniyaadhaan,

Bahaan- naannoo sumaaleetiinii fi

Lixaan- naannoo benishaanguli/gumuz; naannoo gaambeellaa fi sudaaniin kan daangeffammuu dha.

Keewwata kana keewwata xiqqaa kan tumame yoo jiraatellee, daangaawwan oromiya an naannota olla ishe wajjin qabdu, fedhii ummataa buuura godhachuudhaan, naa nnoo dhimmichi ilaaluu wajjiin waliigalteedhaan jijjiiramuun ni danda'a. akka keewwata kana keewwata xiqqaa 2 waliigalteerra ga'amuun yoo dadhabame ak kaataa heera rippaablika dimokraatawaa federaalaa itoophiyaa keewwata 48 tti, mana maree federeeshinichaatiin kan murtaa'u ta'a.

Summary: keewwata kana keewwata xiqqaa kan tumame yoo jiraatellee, daangaawwa n oromiyaan naannota olla ishe wajjin qabdu, fedhii ummataa buuura godhachuud haan, naannoo dhimmichi ilaaluu wajjiin waliigalteedhaan jijjiiramuun ni dand a'a.

Sample 2

Original: alaabaan fi asxaan naannoo oromiyaa eenyummaa uummata naannichaa, tokkummaa, sabboonummaa, gootummaa fi abbaa seenummaa akkasumas walitti hidha miinsa dinagdee kan calaqqisiisu ta'a. alaabaan naannichaa gararraan diimaa, gidduun adii, jalaan gurraacha ta'ee gidduu issaa irratti mallattoo odaa ni q abaata.

Asxaan naannichaa keessa isaatti mallattoo odaa, jalallii qamadii fi giirii warshaa ni qabaata.

Tartiibni alaabaa fi asxaa seeraan murtaa'a.

Summary: alaabaan naannichaa gararraan diimaa, gidduun adii, jalaan gurraacha ta'ee gidduu issaa irratti mallattoo odaa ni qabaata. Asxaan naannichaa keess a isaatti mallattoo odaa, jalallii qamadii fi giirii warshaa ni qabaata.

5.4. Evaluation and Discussion

Rouge (Recall- Oriented Understudy for Gisting Evaluation) evaluation metrics is used to measure the model accuracy. The evaluation is carried out both for abstractive and extractive text summarization methods on same dataset. The results are shown in table 6.4.

Table 6.4: Rouge evaluation results for model accuracy

Method	*Rouge*	*Recall*	*Precision*	*F-measure*
Extractive	Rouge 1	0.511	0.553	0.685
	Rouge 2	**0.521**	**0.662**	0.588
	Rouge L	0.491	0.533	0.485
Abstractive	Rouge 1	0.493	**0.816**	0.683
	Rouge 2	**0.611**	**0.561**	0.473
	Rouge L	0.584	0.693	0.523

Here the ROUGE evaluation results for both the extractive and the abstractive summarization are discussed.

Generally, regarding the relevance of the summary outputs, the abstractive summary output (Precision = 0.816) outperformed the extractive summary output (Precision = 0.662) for ROUGE 1. However, the results of ROUGE 2 indicated that the extractive system summary relevance result (Precision = 0.662) outperformed the extractive system summary (Precision = 0.562).

Regarding the recall values, that is, the extent to which the system summary is captured from the reference summary, showed that abstractive system summary has captured parts of the reference summary (Recall = 0.611) better than that of the extractive system summary (Recall = 0.521) on ROUGE 2. Nevertheless, The recall values by Rouge 1 indicated that extractive system summary (Recall = 0.511) has better captured parts of the reference summary than the abstractive one (Recall = 0.493). When the recall and precision values add considered Abstractive summary output is better than the extractive summary output.

CHAPTER SIX

6. CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORKS

6.1. Conclusions

In this study, Textrank algorithm was employed to implement extractive summarization and Recurrent Neural Networks (LSTM) for abstractive summarization on the Afan Oromo Proclamation texts. ROUGE 1, ROUGE 2 and ROUGE L evaluation metrics were used to evaluate the system summaries. As the results indicate, the abstractive system summary (Recall = 0.611) showed better representation results in capturing the ideas in the reference summary than the extractive system summary (Recall = 0.521) on ROUGE 2 but the extractive system summary (Recall = 0.511) showed a bit better representation over abstractive (Recall = 0.493) In addition, the fact that the results of precision measure indicate that the abstractive summary output (Precision = 0.816) outperformed the extractive summary output (Precision = 0.662) for ROUGE 1 indicated that the abstractive summary is by far better relevant. In addition, the ROUG L recall (0.491) and precision (0.533) values showed better results for the abstractive system summary when compared with that of the extractive recall (0.584) and precision (0.693) values in this case. In this context, the generated abstractive summary is better than the extractive and the RNN (LSTM) algorithm used has delivered better output than the algorithm of extractive summarization in this case. In addition, the result of the graphic representation of training and loss values (Figure 5.2) showed that a good fit of a model or algorithms to summarized dataset is realized on both the extractive and abstractive due to the point of stability occurs with the minimal gap between the final training and validation loss values. Though the system summary of the abstractive method outperformed the extractive one, the graphic representation of the training and loss values indicated that both the algorithms were fitting for the purpose. It is the matter of degree of relevance that puts abstractive summary and its corresponding algorithm at better position.

To conclude, the analysis of the recall and precision values indicate that the generated abstractive system summary of the Afan Oromo Proclamation is by far better and the algorithm used for the Afan Oromo proclamation texts is also fitting. However, the system summary output of the extractive summary is also good and its corresponding algorithm is also of good fit.

Regarding the comparisons results of the reviewed prior works done on Afan Oromo texts, the results of the research on “Afan Oromo Text Summarizer Using Word Embbeding” by Lamesa [10] indicated Recall = 0.422, Precision = 0.527 and F-measure = 0.468 on ROUGE 1 as well as Recall = 0.422, Precision = 0.527, and F-measure = 0.468 for ROUGE 2. These results go in line with the results of the current research work though the current outcome is better. When compared with results obtained on the prior work done by Gammachis [12] (Recall = 0.621, Precision = 0.66 and F-measure = 0.722 on ROUGE 1 and Recall = 0.741, Precision = 0.783 and F-measure = 0.863 on ROUGE 2. This result outperformed the current research result though the corpus used and the methodology employed are not the same.

6.2. Contributions

The contributions of this research work are:

- It attempted to conduct extractive and abstractive text summaries on Afan Oromo proclamation texts as dataset for the first time and developed an insight for the language text document summarization.
- It might be a stepping stone for the future works to be done on the language texts documents.
- Also it might have contributed to the state of the art by recommending future works.

6.3. Future Works

To build better insight on the summarization of the Afan Oromo proclamation texts in particular and that of the Afan Oromo texts in general, as well as to contribute for the advancement of the state of art the following future works might be important.

- Since almost all the prior research works of Afan Oromo Text summarizations are only done on extractive summaries there is a need to undertake abstractive text summarizations.
- Since many of the extractive summaries were done on news text summarizations, there is a dire need to consider other Afan Oromo text datasets on both the extractive and the abstractive summarizations.

- To conduct effective Afan Oromo text summarizations require pre-prepared standard lexicon of the language.

REFERENCES

[1] Dalal, V., & Malik, L. G. (2013, December). “A Survey of Extractive and Abstractive Text Summarization Techniques”. In Emerging Trends in Engineering and Technology (ICETET), 6th International Conference on 2013, pp. 109 -110

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678

[3] Mani, I., *Automatic Summarization.* Amesterdam: John Benjamin's Publishing Company, 2001

[4] Radev, E. Hovy, K. McKeown, "Introduction to the Special Issue on Summarization", *Computational Linguistics*, Vol. 28, No. 4, pp. 399-408, 2002

[5] Ziqiang Cao, Wenjie Li, Sujian Li, and FuruWei. “Improving multi-document summarization via text classification”. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017). San Francisco, United States, 2017.

[6] “The Housing Census: Population Size by Age and Sex, Summary and Statistical Report “Addis Ababa, 2007.

[7] Tabor Wami's, “*Yewugena Dirsetochena Yetarik Ewunetawoch”*, Berhanina Selam Printing Press, Addis Ababa: Ethiopia:, 2004.

[8] Ibrahim Bedane (2015) “The Origin of Afan Oromo: Mother Language”. *Global Journal of HUMAN-SOCIAL SCIENCE: G Linguistics & Education* Volume 15 Issue 12 Version 1.0, USA: Global Journals Inc. 2015

[9] Girma Debela, “Afan Oromo News Text Summarizer”, A thesis presented at Addis Ababa University: Addis Ababa, Ethiopia, 2012

[10] Lamesa Teshome, Afan Oromo Text Summarization Using Word Embedding, A thesis presented at. Addis Ababa University: Addis Ababa, Ethiopia, 2021

[11] Fiseha Berhanu, “Afan Oromo Automatic News Text Summarizer Based on Sentence Selection Function”, A thesis presented at Addis Ababa University: Addis Ababa, Ethiopia, 2013

[12] Gammachiis Temesgen, “Afan Oromo Text Summarization Using Sentence Scoring”, A thesis presented at St. Mary’s University: Addis Ababa, Ethiopia, 2021

[13] Sisay Abera, ‘Information Extraction Model From Afan Oromo News Texts’, A thesis presented at Bahir Dar University: Bahir Dar, Ethiopia, 2020

[14] V. Gupta and G. S. Lehal, “A Survey of Text Summarization Extractive techniques”, *J. Emerge Technol,Web Intell*, vol 2. No.3. Pp. 258-268, 2010

[15] Gu J, Lu Z, Li H, Li VO. “Incorporating copying mechanism in sequence-to-sequence learning”.[Online] Available at: arXiv Prepr arXiv:1603.06393 . 2016.

[16] Shengli Song, Haitao Huang, and Tongxiao Ruan. 2018. “Abstractive text summarization using lstmcnn based deep learning” [Online]. *Multimedia Tools and Applications*, pp857 – 875, 2018 Available at *https://doi.org/10.1007/s11042018-5749-3*

[17] N. R. Kasture, N. Yargal, N. N. Singh, N. Kulkarni, and V. Mathur, “A survey on methods of abstractive text summarization,” *International Journal for Research in Emerging Science and Technology*, vol. 1, no. 6, p. 5, 2014.

[18] M. Allahyari, S. Pouriyeh, M. Assefi et al., “Text summarization techniques: a brief survey,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017.

[19] Sharef, N.M., A.A. Halin and N. Mustapha, “Modelling knowledge summarization by evolving fuzzy grammar”. *American Journal of Applied Science*, 10: 606-614. DOI: 10.3844/ajassp.2013.606.614

[20] Zhang, Y., D. Wang and T. Li, iDVS An interactive multi-document visual summarization system. Mach. Learn. Know. Disco. Databases, 6913: 569-584. DOI: 10.1007/978-3-642-23808-6_37, 2011

[21] Chen, Y. C. Bansal, M, (2018) Fast abstractive summarization with reinforce selected sentence rewriting [Online] Available: *http:*//arXiv preprint arXiv. 1805. 11080

[22] A. B. Al-Saleh and M. E. B. Menai, “Automatic Arabic text summarization: a survey,” *Artificial Intelligence Review*, vol. 45, no. 2, pp. 203–234, 2016.

[23] Q. A. Al-Radaideh and D. Q. Bataineh, “A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms,” *Cognitive Computation*, vol. 10, no. 4, pp. 651–669, 2018.

[24] C. Sunitha, A. Jaya, and A. Ganesh, “A study on abstractive summarization techniques in Indian languages,” *Procedia Computer Science*, vol. 87, pp. 25–31, 2016.

[25] A. Khan and N. Salim, “A review on abstractive summarization methods,” *Journal of @Teoretical and Applied Information Technology*, vol. 59, no. 1, pp. 64–72, 2014.

[26] Y. Jaafar and K. Bouzoubaa, “Towards a new hybrid approach for abstractive summarization,” *Procedia Computer Science*, vol. 142, pp. 286–293, 2018.

[27] Mani, I., "Automatic Summarization" ,John Benjamin's Publishing Company, 2001

[28] Gadaa Malbaa, "Oromia" ,Sudan: Khartoum, 1988

[29] Inderjeet MANI, "Summarization Evaluation: An Overview", The MITRE Corporation, W640 11493 Sunset Hills Road Reston, VA 20190-5214, USA

[30] George Pachantouris, " GreekSum", Master Thesis, DSV,2005

[31] Nima Mazdak; 'A Persian text summarizer". Master Thesis, Stockholm University

[32] Gong, Y. & Liu, X. ,"Generic text summarization using a trainable summarizer and Latent Semantic Analysis", In Proceedings of 24th annual international ACM SIGIR conference on research and development in information retrieval (SGIR'01) (pp. 19-25), New Orleans, LA, USA, 2001.

[33] Ani Nenkova and Kathleen McKeown. "Automatic Summarization' *Foundations and Trends in Information Retrieval* Vol. 5, Nos. 2–3, pp. 103–233, 2011

[34] Hovy, E. H. "Automated Text Summarization". In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583-598. Oxford University Press, 2005.

[35] Sparck Jones, K. "Automatic summarizing: factors and directions". In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999.

[36] Bengio, Y. (2009). "Learning deep architectures for AI". *Foundations and trends® in Machine Learning*, 2(1), 1-127.

[37] Bengio, Y., & Le Cun, Y. (2007). Scaling learning algorithms towards AI. *Large-scale kernel machines*, 34(5).

[38] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798-1828.

[39] Arel, I., Rose, D. C., & Karnowski, T. P. (2010). "Deep machine learning-a new frontier in artificial intelligence research" [research frontier]. *IEEE Computational Intelligence Magazine*, 5(4), 13-18.

[40] Hinton, G. E., Osindero, S., & Teh, Y. W. "A fast learning algorithm for deep belief nets. *Neural computation"*, 18(7), 1527-1554. 2006

[41] Glorot, X., & Bengio, Y. "Understanding the difficulty of training deep feed forward neural networks". Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249–256). 2010

[42] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. "Ranking with Recursive Neural Networks and its Application to Multi-document

Summarization”. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015). Austin, United States

[43] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. [Online], Available at: arXiv preprint arXiv:1412.3555, 2014

[44] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. *Neural Computation,* V. 9, pp, 1735–1780, 1997

[45] Piji Li, Wai Lam, Lidong Bing, Weiwei Guo, and Hang Li.. “Cascaded Attention based Unsupervised Information Distillation for Compressive Summarization”. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017

[46] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. [Online], Available at: arXiv preprint arXiv:1508.01991, 2015

[47] Xin Zheng, Aixin Sun, Jing Li, and Karthik Muthuswamy. “Subtopic-driven Multi-Document Summarization”. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019

[48] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997

[49] K. Al-Sabahi, Z. Zuping, and Y. Kang, *Bidirectional Attentional Encoder-Decoder Model and Bidirectional Beam Search for Abstractive Summarization*, Cornell University, Ithaca, NY, USA, [Online], Available at: http://arxiv.org/abs/1809.06662, 2018

[50] A. M. Rush, S. Chopra, and J. Weston, “A neural attention model for abstractive sentence summarization,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015

[51] K. Lopyrev, Generating news headlines with recurrent neural networks,[Online] Availableat: https://arxiv.org/abs/1512.01712, 2015

[52] Chin-Yew Lin. 2004. “Rouge: A Package for Automatic Evaluation of Summaries”. In Text Summarization Branches Out. 74–81

[53] Afsaneh Rezaei, Sina Dami, and Parisa Daneshjoo, ”Multi-Document Extractive Text Summarization via Deep Learning Approach”, 5th Conference on Knowledge-Based Engineering and Innovation, Iran University of Science and Technology, Tehran, Iran, 2019

[54] Mohammad Masum, Sheiki Abujar, Ashharaful Islam, “Abstractive Method of Text Summarization with Sequence to Sequence RNN, Conference paper, Kanpur: India, Dec. 2019

[55] Abaynew Guadie, Debela Tesfaye, and Teferi Kebebew “Amharic Text Summarization for News Items Posted on Social Media”, Jimma University, Jimma Institute of Technology, Faculty of Computing, Jimma: Ethiopia, 2021

[56] Girma Debele, "Afan Oromo news text summarizer," Unpublished Master's thesis, Addis Ababa University, Ethiopia, 2012

[57] Tilahun G., "Qubee Afan Oromo : Reasons for choosing the Latin script for developing an Afan Oromo Alphabet," *Journal of Oromo studies*, 1993

[58] C. Griefenow-Mewis, W. J.G, Möhlig and B. Heine, "A Grammatical Sketch of Written Oromo," Grammatical Analyses of African Languages, vol. 16, 2001

[59] Debela Tesfaye, "Designing a Stemmer for Afan Oromo Text: A hybrid approach" Unpublished Master‟s thesis, Addis Ababa University, Ethiopia, 2010

[60] Gumii Qormaata “Afan Oromoo, Caasluga Afan Oromo”, Finfinnee: Komishinii Aadaaf Turizmii Oromiyaa, 1995

[61] Nallapati, R., Zhou, B., Santos, C., and Xiang, B. “Abstractive Text Summarization Using Sequence to Sequence RNN and Beyond,” ICLR workshop, abs/1602.06023, 2016

[62] Hobbs, J.”A model for natural language semantics, Part I: The model. Technical report”, London: Yale University, 1974

[63] Magalata Oromia, “Labsi Afan Oromo”[Online] Available: at oag.gov.et and/or www.caffeeoromiyaa.org

[64] R. Mihalcea, and P. Tarau, “TextRank: Bringing Order into Texts, University of North Texas, Department of Computer Science [Online] Available at: {rada,tarau}@cs.unt.edu. 2

[65] R. Darmawan, and R. Satria, “Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization,” *Journal of Intelligent Systems,* 2015, Vol. 1, No

[66] Lin, C. Y. “Looking for a few good metrics: ROUGE and its evaluation”. In Proceedings of NTCTR workshop, Japan: Tokyo, 2004

[67] Afaq, S. and Rao, S. “Significance of Epochs on Training a Network”, [Online] International Journal of Scientific and Technology Research, Vol. 9, 2020, Available at: IJSTR@2020 wwwijstr.org