International Journal on Data Science and Technology
2023; 9(1): 1-12
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20230901.11
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Extractive Text Summarization Using Deep Learning for Tigrigna Language

Meresa Hiluf Gebrehiwot[1], Michael Melese[2]

[1]Faculty of Electrical/Electronics and ICT, Federal Technical and Vocational Education and Training Institute, Addis Ababa, Ethiopia
[2]School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Email address:

Meresa.hiluf@aau.edu.et (Meresa Hiluf Gebrehiwot), Merehil2011@gmail.com (Meresa Hiluf Gebrehiwot), michaelmelese@gmail.com (Michael Melese)

To cite this article:

Meresa Hiluf Gebrehiwot, Michael Melese. Extractive Text Summarization Using Deep Learning for Tigrigna Language. *International Journal on Data Science and Technology*. Vol. 9, No. 1, 2023, pp. 1-12. doi: 10.11648/j.ijdst.20230901.11

Received: October 31, 2022; **Accepted**: January 4, 2023; **Published**: March 20, 2023

Abstract: With the ever-increasing amounts of textual material such as web pages, news articles, blogs, microblogs, and similar, the Internet became the massive body of unstructured information. In this paper to deal with the issues for the availability of more and more information with less time, the extractive text summarization using the deep learning model was used. In this paper, the proposed approach uses three basic stages of feature extraction, feature enhancement, and summary generation of the given news article to extract the core information, to produce well understandable summary and save reader's time. In the feature extraction, We explore various features to improve the extracted sentences to the summary by the score and rank of the extracted features matrix by calculating the top thematic words, paragraph segmentation, sentences length & position, proper nouns, and TF-ISF, and the sum of the feature vector given to RBM to enhance the extracted feature vector and finally generate the final summarization by taking top high scores and 50% 0f the sum second higher scores from the enhanced feature extracted scores. For experimenting purpose, we have used 10 news articles from the total gathered news articles gathered from BBC-Tigrigna, Fana-Tigrigna and VOA-Tigrigna news website. The evaluation of the extracted summary was evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to compare the system extracted summary with the reference / manual summary prepared by human experts. According the experimentation, the average score of ROUG-1 shows 49% for recall, 39% precession, 42% for F-score and for the ROUGE-2 shows that 32% recall, 26% precession and 28% for F-score, for ROUGE-l also shows that 39% of recall, 33% of Precession, and 35% of F-scores. The result shows the proposed approach have higher result in Rouge-1 and the F-score or harmonic mean of precision and recall is 42% and it solves the problems of information overloading in the ever-increasing available news articles by generating the extractive summarizations.

Keywords: Extractive, Deep Learning, Restricted Boltzmann Machine (RBM), Sentence Features, Single Document, Summarization, Unsupervised

1. Introduction

1.1. Background

Text summarization is a way used to reduce the original text data into smaller ones without losing its meaning, and eventually, saves the readers time [2]. With ever-increasing amounts of textual data available in the digital space, such as news articles, blogs, microblogs, and similar with havening less and less time needs to have summarized text data. Now adays, the Internet became the massive body of unstructured information [1]. The Automatic text summarization was assisting us with acquiring relevant information within less time from the available more and more unstructured data/text accessible on the web. For such kind of issues, Natural language Processing (NLP) assumes an essential part in arising an emerging text summarization dependent on the idea of various explicit languages.

There are various classifications of Text Summarization (TS) techniques categorized by various researchers. Those classifications were, based on the information input type as single or multi-document, based on the purpose general or domain-specific, and based on the output type also categorized as extractive or abstractive text summarization [2].

Among these classifications, an extractive summarization method is concatenating important sentences or paragraphs without understanding the meaning of those sentences to produce the subset of the original text document. In the case of an abstractive text summarizing approach, the system must comprehend the meaning of the original text document in order to generate a paraphrased text document with different phrases or sentences but the same meaning to the original document.

In this study, we use an extractive text summarization for unsupervised single document using the Deep Belief Networks (DBNs) composed of stacked layers of Restricted Boltzmann Machines (RBMs) for the Tigrigna language. Because the selected language was morphologically rich. The same word can have multiple meanings and difficult for abstractive text summarization. The Tigrigna language was a Semitic language spoken in the Tigray Region of Northern Ethiopia and Eritrea, according to Abraham Negash [3].

1.2. Problem Statement

With ever-increasing amounts of textual material there is more and more data available in web pages, news articles, blogs, microblogs, and other source of information and have less and less time to get the important information need to have summered text documents. The proposed solution is produced condensed summary, or producing subset of the original documents.

1.3. Objectives

1.3.1. General Objectives

The main objective of the study was to build a text summarizer for Tigrigna news articles by identifying the most important information from the given text and present it to the end users using deep learning neural networks.

1.3.2. Specific Objectives

Specific objectives of the proposal are listed as follow:

1) Reviewing and analyze automatic text summarization methods.
2) Designing and developing an extractive text summarizer for the Tigrigna news article.
3) To produce a condensed summary of news articles.
4) To evaluate the performance of the Tigrigna text summarizer.
5) To report the finding of the study for the upcoming research area.

1.4. Literature Review

In this literature reviews cover the overview the text summarization, type of text summarization, approaches of the text summarization, and the evaluation methods of the text summarization was review in detailed.

Text summarization is the process of making large documents into smaller ones without losing the context, which eventually saves readers time [4]. Automatic Text Summarization is a growing field of study in NLP and becoming a popular/hot research area due to the growth of data and the need to process it more efficiently in the last few years [5]. Automatic text summarization is part of machine learning, natural language processing (NLP), data mining and becoming a popular research area while data grow and there is a demand to process it more efficiently [5].

Generating a summary requires considerable cognitive effort from the summarizer (either a human being or an artificial system): different fragments of a text must be selected, reformulated, and assembled according to their relevance. The coherence of the information included in the summary must also be taken into account [6]. Natural language processing (NLP) plays an important role in developing an automatic text summarization based on the nature of different specific languages [4]. This can be done using different techniques like TextRank using a graph-based ranking algorithm, Feature-based text summarization, LexRank using TF-IDF with a graph-based algorithm, Topic-based, using sentence embeddings, and for deep learning techniques using word2vec and Encoder-Decoder Model.

The first applications in history of text summarization were library catalogs in 1674 and later generating abstracts for research articles in 1898 [7]. At first, the emphasis was on generating summaries that would help choose the best articles for deeper reading rather than generating summaries that would replace the original text.

The first summarization system was built on the first commercial computer, IBM 701, by Luhn in the 1950s and it was based on bag of words technique and counting word frequencies. He extracted frequently occurring words and then gave each sentence a number based on how frequent words the sentence has. The number presented the significance of the sentence. Then the abstract was formed of the most significant sentences. [8]

A decade later Edmundson (1969) [9] introduced new statistical methods on automatic extraction based on Cue, Key, Title, and Location methods. The Cue method aims to have a corpus of words whose appearance in a sentence would make the sentence either important, unimportant or irrelevant. The Key method selects the words that appear in the original text more frequently than in the whole corpus being the start for the TF–IDF (term frequency-inverse document frequency) method, the Title method takes into account the title and the headings, and the Location method is the position of the sentences: sentences under headings of first and last sentences of paragraphs and the document are usually more relevant than other sentences. He also emphasized that semantic and syntactic features of the text should be taken into account in the future development of summarizers, e.g., the length of the summary could be determined automatically, Edmundson set it to 25% of the sentences in the original.

Little by little linguistics was taken into account and systems started to handle different word forms with the techniques of NLP. The focus was on extracting, categorizing, and classifying text. Between 1990 and 2000 machine learning was introduced in NLP to parse sentences into tokens and stemming words into their base forms [7].

So far, the research focused solely on words, and computers were not able to understand the semantics of a text document. Text analytics was anyway evolving rapidly in the next phase intending to move to understand the meaning of the text occurs. researchers are still trying to build systems that can gently understand the semantics and pass reading comprehension tests.

Based on [10-13], RBM's have traditionally been used in computer vision tasks. However, these recent works have shown that they can be very effective in Natural Language Processing (NLP) tasks as well. The specific model they implemented was a regression process for sentence ranking. The architecture of this method consists of a convolution layer followed by a max-pooling layer, on top of a pre-trained word2vec mapping. Leo Laugier *et al* [10] implemented the proposed method and perform experiments on single document extractive summarization. They show that RBM can achieve superior performance than state-of-the-art systems. They use Python as a tool because python has versatility, the capability of fast production, and it has great support from a deep learning framework [10]. They also use the evaluation metrics of ROUGE-1 and ROUGE-2 to compare the results for the dataset of Document Understanding Conferences (DUC) from 2001 to 2004. Rouge accesses the quality of an automatic summary by counting the overlapping units, such as n-gram, common word pairs, and longest common sub-sequence between automatic summary and a set of reference summaries [12].

The development of text summarization was developed ever increased through the availability of more and more unstructured data and the need to process an efficient way of text analysis. Text summarization was developed from the stoical and machine learning-based summarization to the NLP familiar methods like using deep learning approaches. Our main target was to analyze, how to produce extractive text summarization using the Deep belief networks stakes over Restricted Boltzmann Machine, which was, one of the algorithms of deep learning for the local languages like Tigrigna. Additionally, we make available the source for others as basement like the dataset we prepare.

1.5. Related Works

The related works in this paper covers the summarization technique, document size, summary type, and approach for Tigrigna langue summarized on the following table.

Table 1. *Related Works.*

Author	Title	Purpose/Objective	Methodology	Key Finding/result	Gaps
Guesh Amiha (2017)	Automatic Text Summarizer for Tigrinya Language	1) To produce 'title selection' using 'term frequency 'and content summary-based thematic generic words.	1) 'term frequency and title word' statistical algorithm using the tool python	1) For the two features of term frequency and title word, he obtained as 58% and 60% respectively.	1) dimension of the inputs was not clearly defined either, single-document or multiple documents
Mulugeta Getachew (2017)	Topic-based Tigrigna Text Summarization Using WordNet	2) To produce Tigrigna news article using semantic similarity and topic from a single text document	2) Uses Probabilistic Latent Semantic Analysis (PLSA) over the WordNet 'semantic similarity and topic of a document	2) Scoring method for top 15 of 20 restarts gives the best-summarized document for 25% extraction rate which was 48.5%.	2) There is not available adequate well-organized wordnet for the Tigrigna language.
Melese Tamiru (2009)	Automatic Amharic Text Summarization Using Latent Semantic Analysis	3) To produce generic text summarizations using the methods of ranking and extracting sentences from the original documents.	3) He uses latent semantic analysis (LSA) to identify the main topics of a document and 4) LSAGraph, combines LSA with graph-based ranking algorithms	3) He uses F-score for 20% and 30% of extraction for both methods and 4) He found the result of 30% extraction have better result 0.47 for TopicLSA and LSAGraph + PageRank 0.45, LSAGraph + HITS 0.47.	3) He recommends developing a sentence compression algorithm for Amharic
Eyob Delele Yirdaw (2011)	Topic-based Amharic Text Summarization	4) To develop concept-based single-document Amharic text summarization system	5) He uses statistical approaches called topic modeling. 6) topic modeling approach of probabilistic latent semantic analysis (PLSA)	5) The result was evaluated using precision/recall for summaries of 20%, 25%, and 30% extraction rates. 6) The best results achieved as 0.45511 at 20%, 0.48499 at 25%, and 0.52012 at 30%.	4) Recommends applying for multi-document, query-focused, and update summarization. 5) He identifies inflectional morphology only.
Addis Ashagre Teklewold (2013)	Language independent single document text summarization using C#	5) To produce language-independent single document text summarization	7) Combinations of term frequency and sentence position methods were used to rank the sentences	7) F-measure comparing them with an ideal manual summary using news article extraction rates at 10%, 20% and 30%. 8) The highest score is 75.65% at the 30% and the learning-based has been achieved.	6) Recommends, multi-document text summarization for Amharic and other languages.
Mattias Gessesse	Efficient Language-	6) To produce language-independent text	8) Uses two algorithms called Independent Rank	9) The algorithm was tested for English and Amharic using	7) This is also, focuses on the statistical frequency

Author	Title	Purpose/Objective	Methodology	Key Finding/result	Gaps
Argaw (2015)	Independent Text Summarization Using Graph-Based Approach	summarizations	(IR) and Sentence Rank (SR) rather than existing adopted TextRank and LexRank used for page rank and rank the numbers of links	ROUGE-1 result of 0.5238 on half of 2002 DUC dataset for stemmed and the highest of them reporting 0.4405 ROUGE-1 results and 10) TextRank, which uses a graph-based approach, with a reported ROUGE-1 result of 0.4229 on the same data set.	of the give text document.
Mohammed Abdella Hassen in 2016	automatic Amharic text summarizer using an abstractive approach	7) To produce abstractive text summarization for the Amharic language.	9) Uses Universal Networking Language (UNL) which is one of the semantic representations of natural language sentences	11) evaluation is promising since they use the subjective evaluation of the summary sentence	8) Recommends that the Machine learning approach can be considered for the processes of UNL related tasks.
Abaynew Guadie, Debela Tesfaye, Teferi Kebebew in 2021	Amharic Text Summarization for News Items Posted on Social Media	8) To summarize the news items posted Amharic texts posted in twitter and face book	10) extractive summarization approach by Calculating similarity of each posted doc, with two pair of sentences and cluster-based similarity using Kmeans algorithm	12) the highest F-measure score is 87.07% for extraction rate at 30%, in the clustered group of protests posts. 13) 2nd highest F-measure score is 84% for extraction rate at 30%, in droughts post groups 14) 3rd highest F-measure score is 91.37% for extraction rate at 30%, in the sports post groups and 15) 4th highest F-measure score is 93.52% for extraction rate at 30% to generate the summary post texts	9) It not clearly stated why the clusterin Kmean algorithm was used 10) They recommend to identify the texts posted in social media automatically rather than use manually, apply more Amharic lexicon rules and dictionaries files to use and control over and underestimations

2. Research Methodology

In these sections, the proposed research design and methodology for the selections of research type, approach, data preprocessing and production of the summary was explained in details.

2.1. Research Design

In this paper, we used Design Science research methodology to apply extractive text summarization for Tigrigna news article. The design science research involves the construction and evaluation of Information Technology artifacts, constructs, models, methods, and instantiations [14]. Design science could be a problem-solving paradigm that looks to enhance human knowledge through the creation of innovative artifacts. DSR seeks to enhance technology and science knowledge bases through the creation of innovative artifacts that solve problems and improve in the environment in which they are instantiated [15]. The results of DSR incorporate both the recently designed artifacts and design knowledge (DK) that gives a fuller understanding by means of design theories of why the artifacts enhance the significant application contexts [14, 15]. In this case, Design Science research method was used to design and construct extractive text summarization using the architecture designed below in Figure 1.

In this study, we follow all the stages of design science process that includes, the problem identification and motivation, solution objectives, design and development, evaluation, and communication.

2.2. Data Set

In this paper, we prepare the dataset from the available news article for Tigrigna language in Voice of America (VOA) Tigrigna, Fana Broad Casting (FBC) Tigrigna, Dmtsi Woyane Tigray and BBC Tigrigna news. For the exploratory purpose, each of those articles was free from tables and figures within the report sources. The taking after tables appears the details of our information set.

Table 2. *Statistics of the Data set.*

Data Set Attribute	values
Total numbers of news articles	1900
Min sentence per news article	103
Max sentences pre news article	10
Average sentences per news article	25
Max numbers of words	1990
Min numbers of words	142
Average words	434

3. Data Analysis, Architecture, or Experimentation

3.1. Architectures

The architectures of the selected model show all the stages of extractive text summarization. The primary goal of this system is used to select the most frequent words and which sentence should be included in the summary. Figure 1 shows the architecture of our system that contains three phases.

(1) preprocessing: This module consists of four components: text segmentation, tokenization, stop word elimination, stemming, and normalization, and their purpose is to efficiently represent the input text in a suitable format for the subsequent text summarization feature extraction process while maintaining the consistency of the summary.

(2) Feature extraction: After Text Preprocessing, the sentence features are calculated based on their respective formulas given per feature, to get the sentence score. The sentence feature contributes to choosing the sentence score and includes the Number of thematic words, Sentence position, Sentence length, Sentence position relative to paragraph, Number of proper nouns, Number of numerals, Number of named entities, Term Frequency-Inverse Sentence Frequency (TF-ISF), Sentence to Centroid similarity. The scoring of those features dealt with the term's individual score as well as the sentence that included the term. It assigns a score to words that occur in multiple sentences throughout the entire text.

(3) Feature Enhancement**:** Following the extraction of those nine features, the features are augmented using RBM depending on their scores. We combine and convert the present features in the datasets into a smaller collection of features that we can utilize for summarization, clustering, classification, and other tasks when we undertake feature extraction. This was done to reduce overfitting and get better outcomes in less time. Each phrase comprises 9 feature vector values, which were used to construct the sentence-feature matrix. The feature vectors are then enhanced and abstracted, allowing complex features to be built out of simple ones. The sentence-feature matrix is fed into a Restricted Boltzmann Machine (RBM) with one hidden layer and one visible layer to improve those features. This step enhances the summary's quality.

(4) Summary generation: Using the selected deep learning model, this module is responsible for determining the best candidate sentences for the summary.

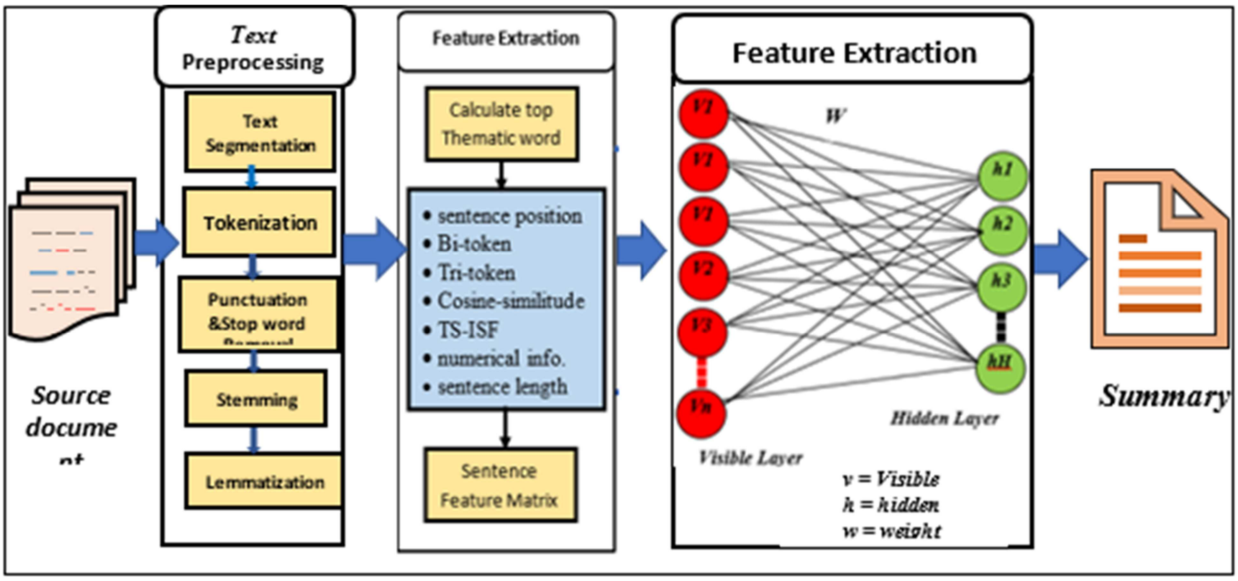

Figure 1. *General Architecture of extractive text summarization using deep Learning* [16].

3.2. The Proposed Approach

This paper goes through unsupervised extractive text summarization by utilizing deep learning approaches (i.e., Restricted Boltzmann Machine (RBM)) for single-document summarization. This was applied in to three phases of feature extraction, feature enhancement and summary generation based on scored values of those features [16-18]. Those phase work together for the purpose of integrating the main information combined in each phase and generate a summary.

Based on this, the sentences that contain the thematic words are scored using the sentence-feature matrix. The primary goal of this system is selecting the most frequent words and which sentence should be included in the summary. Figure 1 shows the architecture of our system that contains three phases. (1) pre-processing: this module consists of four components: text segmentation, tokenization, normalization, stop word removal, and stemming, and their purpose is to efficiently represent the input text in a suitable format for the subsequent text summarization process while maintaining the consistency of the summary. (2) Feature Extraction: this module dealt with the term's individual feature extractions based on the score of the given 9 features of the number of thematic words, Sentence position, Sentence length, Sentence position relative to paragraph, Number of proper nouns, Number of numerals, Number of named entities, Term Frequency-Inverse Sentence Frequency (TF-ISF), Sentence to Centroid similarity. (3) Feature Enhancement: The feature vectors are then enhanced and abstracted, allowing complex features to be built out of simple ones. The sentence-feature matrix is fed into a Restricted Boltzmann Machine (RBM) with one hidden layer and one visible layer to improve those features. sentences are graded in this module based on the intersection of the most frequent terms. Sentence that contains the most frequent words should have ranked first. (4) Summary generation: this module is in charge of choosing the best candidate sentences for the summary using the selected deep learning model and fuzzy logic section.

3.2.1. Feature Extraction

The text is built into a sentence-feature matrix, once the ambiguity has been minimized and ambiguities have been eliminated. The sentences-feature vector is created for every sentence in the text. The framework is made up of these feature vectors. We've tried out a few different features. The sentence features like the number of the thematic words, sentences position, sentences length, sentences position relative to paragraph, numbers of proper nouns, number of numerals, number of named entities ant the Term Frequency-Invers Sentences Frequency (TF ISF) have proven to be the most effective at summarizing accurate studies [17]. These calculations are carried out on the text that has been obtained following the preprocessing phase:

I. Number of thematic words: The thematic words were taken from the upper10 most frequently occurring words of the sentences. For each sentence, the proportion of numbers of thematic words to total words was determined.

Topics are themselves noun phrases, which we distinguish and extract dependent on part of speech patterns. At that point we score the relevance of these expected subjects through an interaction called lexical chaining. Lexical chaining is a low-level text analytics measure that interface sentences by means of related nouns.

Whenever we've scored the lexical chains, themes that have a place with the highest scoring chains are appointed the highest relevancy scores. We are ready to see that those themes work effectively in passing notifiable information on the context of the article. What's more, scoring these Themes dependent on their context-oriented significance helps us see what's truly significant. themes scores are especially helpful in contrasting numerous articles across time with recognize patterns and trends. The significance of Theme Extraction and Scoring is to Limits to phrases that coordinate certain part-of-speech patterns, score based on contextual pertinence and importance, and Includes sentiment scores for themes.

$$Sentences\ Thematic = \frac{No.of\ thematic\ words}{Total\ words} \quad (1)$$

II. Sentence position: This feature is calculated the first function and returns 1 if the position of the sentence is within a given first or last sentence of the text and cos ((SenPos – min) ((1/max)-min)) otherwise; calculates as follows.

$$Sentence\ Position = \begin{cases} 1;\ if\ its\ the\ first\ or\ last\ sentence\ of\ the\ text \\ cos((SenPos - min)((1/max) - min));\ otherwise \end{cases} \quad (2)$$

where, *SenPos* = position of sentence in the text

min = th x N

max = th x 2 x N

N is total number of sentences in document

th is threshold calculated as 0.2 x N

By this, we get a high feature value towards the beginning and ending of the document, and a progressively decremented value towards the middle.

III. Sentence length: This element is utilized to avoid sentences that are too short as those sentences won't pass on much information. The first function is 0, if the number of words is less 3 and numbers of words in the sentences, otherwise.

$$Sentence\ Length = \begin{cases} 0;\ if\ number\ of\ words\ is\ less\ than\ 3 \\ No{:}\ of\ words\ in\ the\ sentence;\ otherwise \end{cases} \quad (3)$$

IV. Sentence position relative to paragraph: This comes straightforwardly from the perception that toward the beginning of each paragraph, new discussion is started and toward the end of each paragraph, we have a conclusive closing. The function is 1, if it is the first or last sentences of a paragraph and 0 otherwise.

$$Sentence\ Length = \begin{cases} 1;\ if\ it\ is\ the\ first\ or\ last\ sentence\ of\ a\ paragraph \\ 0;\ otherwise \end{cases} \quad (4)$$

V. Number of proper nouns: This feature is used to give importance to sentences having a substantial number of proper nouns. Proper nouns identify people, places or things distinguish explicit individuals, spots, and things. Extracting elements, for example, the proper nouns, places or things make it simpler to mine information. For example, we can perform named entity extraction, where an algorithm takes a string of text (sentence or paragraph) as input and identify the significant nouns (people, place, and organizations) present in it. Here, we count the total number of words that have been PoS tagged as proper nouns for each sentence.

VI. Number of numerals: Since figures are consistently vital to introducing realities, this feature offers significance to sentences having certain figures. For each sentence we calculate the proportion of numerals to total number of words in the sentence.

$$\text{Sentence Numerals} = \frac{No.of\ Numerals}{Total\ words} \quad (5)$$

VII. Number of named entities: Named entity also called element entity identification or entity extraction – is a natural language processing (NLP) method that mathematically distinguishes named entity in a text and groups them into predefined classifications. Entities can be names of individuals, associations, areas, times, amounts, financial values, rates, and more. Here, we count the total number of named entities in each sentence. Sentences having references to named entities like a company, a group of people etc. are often quite important to make any sense of a factual report.

VIII. Term Frequency-Inverse Sentence Frequency (TF ISF): Since we are working with a single news article, we have considered TF-ISF features in to account rather than TF-IDF. Frequency of each word in a specific sentence is multiplied by the total number of occurrences of that word in the wide range of various sentences. We calculate this product and add it over all words.

$$\text{Sentence Numerals} = \frac{\log(\sum_{all\,words} TF*ISF}{Total\ words} \quad (6)$$

IX. Sentence to Centroid similarity: Sentence having the highest frequency of TF-ISF score is considered as the centroid sentence. At that point, we calculate cosine similarity of each sentence with that centroid sentence.

Sentence Similarity = cosine sim (sentence; centroid) (7)

At the end of this phase, we have a sentence-feature matrix.

3.2.2. Feature Enhancement Using RBM

The sentence feature matrix has been generated with every sentence having nine features (Number of thematic words, Sentence position, Sentence length, Sentence length, Number of proper nouns, Number of numerals, Number of named entities, erm Frequency-Inverse Sentence Frequency (TF-ISF), Sentence to Centroid similarity) function vector values. After this, recalculation is carried out in this matrix to enhance and abstract the feature vectors, to construct complicated functions out of easy ones. This step improves the quality of the summary. To enhance and abstract, the sentence feature matrix is given as input to a Restricted Boltzmann Machine (RBM) which has one hidden layer and one visible layer. A single hidden layer was enough for the learning process on the dimensions of the training data.

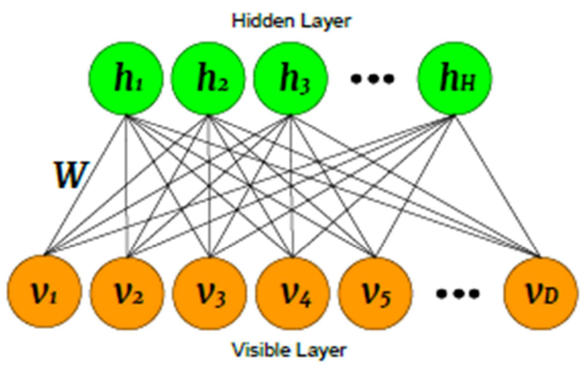

Figure 2. *RBM Architecture based on Nidhi, et al.* [17].

$$E(v,h) = -\sum_{ij,j} W_{ij} v_i h_j - \sum_i v_i b_i - \sum_j h_j c_j$$

Figure 3. *RBM Algorithm.*

FOR EACH iteration DO
Sample a vector v from the data set
SET $e = \sigma(\beta(W^T v + c))$
FOR EACH hidden unit DO
SET $h_j = 1$ with probability e_j
FOR EACH visible unit DO
SET $\bar{v}_i = 1$ with probability $\sigma(\beta(Wh + b))_i$
SET $\bar{e} = \sigma(\beta(W^T \bar{v} + c))$
SET $W = W + \lambda\beta\,[ve^T - \bar{v}\bar{e}^T]$
SET $b = b + \lambda\beta\,[v - \bar{v}]$
SET $c = c + \lambda\beta\,[e - \bar{e}]$
DONE

Figure 4. *One-stem contrastive divergence CD.*

The RBM that we're using has nine perceptron in every layer with a learning rate of 0.1. We use Persistent Contrastive Divergence approach to sample throughout the learning process.

We have trained the RBM for five epochs with a batch length of four and four parallel Gibbs Chains, used for sampling the using Persistent Contrastive Divergence (CD) approach. Every sentence feature vector is going through the hidden layer where in feature vector values for each sentence are multiplied by learned weights and a bias value is included to all of the feature vector values which is also learned by the RBM. At the end, we have a refined and improved matrix. Note that the RBM was trained for every new document that needs to be summarized. The concept was, that no document can be summarized without going over it. Since every file is particular inside the capabilities extracted in section 3.5 above, the RBM was freshly trained for every new document.

```
Procedure
   Initialize the weight matrix W, bias vectors
   a and b, momentum v.
   Set the states of visible unit v1 as the
   training vector
   While i < Max_Iter
     For j = 1, 2 …, m (all hidden units)
       Compute P(h1j = 1|v1) using equation (7)
       Gibbs Sampling h1j ∈ {0,1} from P(h1j|v1)
     End For
     For i = 1, 2 …, n (all visible units)
       Compute P(v2i = 1|h1) using equation (8)
       Gibbs Sampling v2i ∈ {0,1} from P(v2i|h1)
     End For
     For j = 1, 2 …, m (all hidden units)
       Compute P(h2j = 1|v2) using equation (7)
     End For
       //Update rule:
       W := W + ∈ (P(h1 = 1|v1) v1^T − P(h2 = 1|v2)v2^T)
       a := a + ∈ (v1 − v2)
       b := b + ∈ (P(h1 = 1|v1) − P(h2 = 1|v2))
       x := updation of momentum;
   End While
End Procedure
```

Figure 5. *RBM's fast learning algorithm based on CD* [19].

3.2.3. Summary Generation

The Summary generation in RBM, the amended feature vector values are summed to generate a score against individual sentences. The sentences are then sorted according to decreasing score values. In this stored list the furthermost relevant sentences are scored the first sentences and the scored subset of sentences are chosen that forms the summary. At that point the next sentence we select, the sentence having the highest measure of common proximity similarity coefficient with the first sentence, and select strictly from the top half of the sorted list. This handle, incrementally and recursively rehashed to choose more sentences until a user-specified summary constrain is come to. At that point, the sentences are re-arranged within the order of appearance within the original document. This produces a coherent outline instead of a set of confounded sentences.

Summary in Figure 6 shows the approach of creating the summary. The sum of all enhanced feature values for each sentence in the document is calculated and stored in a list.

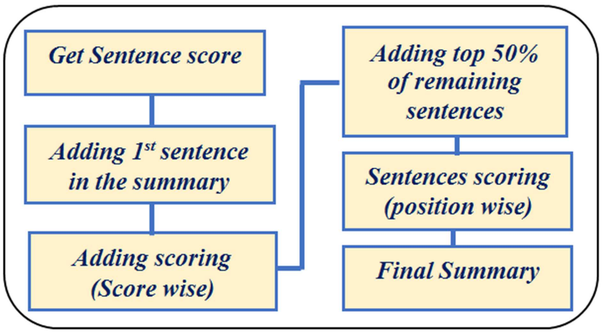

Figure 6. *Generating the Summary* [17].

As a result, a value is generated for each sentence that represents its score. Sentences are ranked in decreasing order based on scores. The first sentence is always included in the summary because it is the most crucial one. Then, the top 50% of the remaining sentences are included and arranged in the first summary in accordance with their original positions in the document.

3.2.4. Evaluation Criteria

To evaluate quality of the system extracted summary adjacent to the human or manual extracted summary using the ROUGE (Recall Oriented Understudy for Gisting Evaluation) evaluation metrics for the recall, precession and F-score of the system summary. For the manual summary, 10 news article was prepared for testing and used as reference to evaluate summary quality to adjacent system summary.

The evaluation criteria in this case, system-generated summary and reference human summary were compared based on three basic measures of Precision, Recall and F-Measure [17]. In terms of precision, we were primarily interested in determining how much of the system summary was actually useful or required. Precision is defined by equation (1) as the following.

$$Precision = \frac{number\ of\ overlapping\ words}{total\ numbers\ of\ words\ in\ system\ summary} \tag{8}$$

In the context of ROUGE, recall refers to how much of the reference summary is recovered or captured by the system summary. By consider the individual words overlapped, it can be calculated as follows (2):

$$Recall = \frac{number\ of\ overlapping\ words}{total\ numbers\ of\ words\ in\ rfernce(manual\ summary)} \tag{9}$$

The F1-Score is represented by the association of Recall and Precision. Accurateness of the result is evaluated by the number of true negative cases, but we focus on false positive and false negative cases instead. Thus, the F1 score is formulated by the following formula:

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall} \tag{10}$$

3.2.5. The Feature Used for Tigrigna Language Text Summarization

The Tigrigna text summary features sets are different from other languages by reading and encoding Unicode files, reading and encoding stop words, punctuation marks, and morphological analyses. In this instance, the module 'codecs. open ()' was used as ‘text = codecs. Open (root.filename, encoding='utf-8'). read ()’, to open and encode the Tigrigna news article to Unicode format.

The focused features of Tigrigna language were typically the stop words and punctuation marks used to analyze and filter the important words out of the given sentences/document. Instead of using a period (.) to divide sentences in Tigrigna, we used fourPoint/Ariba'ete Netibi (።) in the sentence segmentation, and the additional punctuation marks used in Tigrigna language was displayed in the following table.

Table 3. *List of punctuation marks in Tigrinya language* [20].

Punctuation marks	Meaning
። Full stop/period	End of sentences
፡	Word separator
፣ comma	Sentence connector
፤ semicolon	List separator marks
: - Preface colon	Beginning of the list mark
፥ colon	Word separator
? question mark	End of question
! Exclamation	End of n emphatic declaration, or command

The typical stop words that mislead readers are eliminated from the offered articles, and we sort them out according to the frequency of each term in the chosen language.

The common Tigrigna stop words are ['"\”$','\r\nካብ','\r\nምስ','\r\n','ኢና','’የ','’ያ','’ዩ','ኣብ','ንሱ','እውን','ኣብ' ,'እቶም','“ናብ','ክሳብ','ኣብቲ','እቲ','ነቲ','ምስ','ከም','ዘሎ','ብኡኡ','ከምዘሎ',' ከኣ','ውን','ክሳብ','“','ንናይ','ናይ','ናይቲ','እየ','እያ','ኢሉ’ሎ','ኢሉ’ሎ','ምስቲ', ”', 'ካብ', 'እቲ', 'እከለን', 'እወ', 'እዩ', 'እዪ', 'እየን', 'እየ', 'ኢያ', 'ኢሎም', 'ካብ', 'ዘሎ', 'ነታ', "’ያ", 'ዘላ', 'ዘለወን', 'ዘለዋ', 'ኣለዎ', 'ኣብታ', 'ናብቲ','ዘሎ','ኣሎ','ስለ','ስለ','እዚ','እዉን','ኣላ', 'ወን', 'ወይ', 'ኢሉ', 'ግዜ', 'ምስቶም', 'ኢሉ', 'ከም', 'ነዚ', 'ድማ', 'ናብ', 'ኢና', 'ስለዚ', 'ኢለ', 'ናይዚ', 'ናብዚ', 'እዮም', 'ከዓ', 'ኩሉ', 'ከሎ', 'ካዓ', 'ኣብዚ', 'ግን', 'ግና', 'ነዚ', 'ኣቢሉ','እታ','ነታ'].

3.3. Experiments

As discussed in the earlier section, the approaches we used consists of the three stages, feature extraction, feature enhancement, and summary generation, which work together to extract the core information and generate a logical & consistent, understandable summary.

The single Tigrigna news article was selected from the prepared articles randomly as input. Text preprocessing activity was done like splitting paragraph to sentences by separator of ‘arat Netibi ‘(።). After that, sentence was tokenized to words by space delimiter and punctuations, stope words of the article for Tigrigna were removed. We discovered several features to improve the set of sentences selected for the summary and feature matrix base on the position of sentences, bi-token length, tri-token length, TF-ISF feature and Centroid Calculation, cosine similarity, thematic number, sentences length, numeric token, pronoun score as discussed in section 3.2.1. The Restricted Boltzmann Machine (RBM) deep leering algorithm was used to improve resultant accuracy without losing any important information of the above extracted feature matrix. The output of the final summary for the given sample input news article was generated as follow.

Sample input news article


ሞት ጸሊም ብጻዕዳን ጸሊምን እንታይ እዩ ፍልልዩ?
17 ጁን 2020

Figure 7. Sample input News Article for Summary.

Output of System Summary

System_Summary: ጆርጅ ፍሎይድ ሂዩስቴን ቴክሳስ ተቐቢሩ። ልዕሊኡ ዝተፈጸመ ዘስካሕክሕ ጭካነ ሓቢሩዎ ኣይተቐብረን። ጨካን ተግባር ኣእምሮ ኩሉ'ቲ ፍ
ጻሜ ብተንቀሳቓሲ ስእሊ ዝረኣየ ዘበለ ተሓቲሙ ኣሎ። ንጆርጅ ፍሎይድ ክሳዱ ረጊጹ ዝቐተሎ ጨካን ፖሊስ እውን፡ ሕብሪ ቆርበቱን መልክዑን ኣእምሮ ሰባት
ኣይክእለን እዩ። ገበን ጨካን ፖሊስን ምኽንያት ጭካኔኡ ንዝግለጽ ሲስተማቲክ ፖሊሳዊ ምሕደራን እምበር ንኹሎም ጸዓዱ ዝውክል ኣይኮነን። ቅትለት ጆርጅ
ፍሎይድ፡ ኣመሪካን ዝተፈላለየ ክፍልታት ዓለምን፡ ልዕሊ ጸለምቲ፡ ንዘበናት ዝሰረተ ምምሕዳራዊ ስርዓት ዓሌታዊ ኣድልዎን ወጽዓን ንከብቅዕ ዝጽውዕ ሰፊሕ ህዝ
ባዊ ምንቅስቓስ ኣለዓዒሉ'ዩ። ኣስታት ስድራ-ቤትን መቕርብን ተጣበቕቲ ሰብኣዊ መሰላትን ዝተረኽቡሉ ስነ-ስርዓት ቀብሪ "ጆርጅ ፍሎይድ ጸሊም ብምዃኑ ገበ
ነኛ ተቐቲሉ" ዝብል ጽሑፍ ተጠቂዑ ነይሩ። ጭርሖ ቀብሩ ጥራሕ ኣይተነግረን። ብነድሪ ቅትለት ጆርጅ ዝወጽአ "ህይወት ጸሊም ክቡር እዩ!" ዝብል ሰልፍታ
ት ዓለማዊ ጭርሖ'ዩ ክቃላሕ ቀንዩ'ሎ። ብ ዝሓለፈ ሳልስቲ ናይሮቢ፡ ብኣማኢት ዝቑጸሩ ኬንያውያን ሰልፈኛታት፡ መወዳእታ መጋቢት ሕጂ እዋን ዕጽዋ ኮቪድ
-፡ ልዕሊ ሰላማውያን ሰባት ብፖሊስ ሃገር ንዘጋጠመ ቅትለት ዜጋታት ብምዅናን፡ "ዳግማይ ከይድገም" ብዝብል ጭርሖ፡ ፍትሒ ክረጋገጽ ዝጠልብ ሰፊሕ ሰለማ
ዊ ሰልፊ ኣካይዶም። ብዙሓት ካብቶም ሰልፈኛታት ኣስማት ናይቶም ብፖሊስ ዝተቐትሉ ግዳያት ሒዞም ይረኣዩ። ካልኣት ህይወትና ትርጉም የብሉን - መጸኢ
ና ይወሓስ - ዳግማይ ከይድገም" ዝብል ጭርሖ ዝሓዙ ነይሮም። ዝሓለፈ መጋቢት ዝተኣወጀ፡ ምሸት ወጋሕታ ምንቅስቓስ ዝገትእ ሰዓታት እቶ እቶ፡ ዘይሻራዊ
በዓል ስልጣን ምቁጽጻር ፖሊስ ኬንያ ብፖሊስ ተበዲልና ዝብል ጥርዓን ከምዝመጽኦን፡ መስርሕ ከምዝሞቱን ልዕሊ ጉድኣት ከምዝበጽሖምን ሓቢሩ። ጥሕሰት ሕ
ጊ ፖሊስ ኢትዮጵያ ኢትዮጵያ እውን፡ ዝተፈላለየ እዋን፡ ትማልን ሎምን፡ ሓይልታት ጸጥታን ፖሊስን ልዕሊ ሰለማዊ ሰባት ዝተፈላለየ በደላት ፈጺሞምን ይፍጽ
ሙን ኣለው። መወዳእታ ግንቦት ዓመት እኳ ኣምነስቲ ኢንተርናሺናል፡ኦሮምያን፡ ዘስካሕክሕ ግህሰት ሰብኣዊ መሰላት ምፍጻሞም ሰኒዱ ኣሎ። ድሕሪ ጸብጻብ ኣ
ምነስቲ ቀዳማይ ሚኒስተር ኣብዪ ኣሕመድ ፌስቡክን ትዊተርን "ንዕብየት ኢትዮጵያ ዘይሕግዙ ዛዕባታት እዝኒ ኣይንሀቦምን ኢና" ክብል ምጽሓፉ ንብዙሓት ሰ
ባት ኣየሐጎሰን። መንግስታትን እተን ግህሰት ተፈጺሙወን ዝተባህላ ዞባታት ኦሮምያን ኣምሓራን ጸብጻብ ኣይተቐበሉዎን። ይኹን እምበር እዋን ዝተፈለላየ ክፍ
ልታት ኢትዮጵያ መዓልታዊ ሞትን ምምዝባልን የጋጥም ምህላው ብማዕከናት ዜና ብቐጻሊ ይንገር - ኣምነስቲ ኢንተርናሺናል ንነብሲ ወከፍ ፍጻሜ ብዕለቱን ሰ
ዓቱን ጸብጺቡ ኣሰፊሩዎ ኣሎ። ዝሓለፈ ወርሒ ሓደ መንእሰይ፡ ከተማ መቐለ ብልሙድ ' ማይ ዱባ' ዝበሃል ከባቢ ከምእተቐትለን ክልተ ቆሲሎም ንሕክምና ከ
ምዝኸዱን ቢቢሲ ትግርኛ ተጸብጺቡ ነይሩ። ዝኾነ ይኹን ክፍለ ዓለም ሰላማውያን ሰባት ብፖሊስ ተቐቲሎም ክትሰምዕ ዘሰምብድ እዩ። ስለምንታይ ሕጊ ከኽብሩ
ዝግበኦም ፖሊስ ሕጊ ዝጥሕሱ እሞ ኸኣ ሰባት ዝቐትሉ?ኣማራጺ ዝስእኑሉን ህይወት ብዙሓት ንምድሓን ሓይሊ ክጥቀም እንተዘይተገዲዱ፡ ዝኾነ ዝፍጥር ዕግርግ
ር ንህዝቢ ክከላኸለን ክረጋግእን እምበር ንህዝቢ በባትር ኮነ ብተተካሲ ክቐትል ኣይግባእን'ዩ ይብል ጸብጻብ ኣምኒስት ኢንተርናሺናል። ዛዕባ፡ ብዛዕባ ፖሊስ ዝ
ጥቀምዎ ሕሉፍ ሓይሊ ስለዝዛረብ እምበር ካልእ ዝዓበየ ጸገም ፖሊስ ኬንያ ጉቦን ብልሽውንና እዩ። ፖሊስ ሃገር ዝኾነ ነደናታት ከተማ ብፍላይ ንወጻእተኛኣተት
ደው ኣቢሎም ወረቐት መንነት ይሓቱ እሞ ዋላ ጽሩይን ሕጋውን ሰነድ እናሃለዎ ዝሕተት ሰብ፡ ' እንዳ ፖሊስ ኬድካ ኮለል ትብል ገንዘብ ከፈልካ እንተተገላገል
ካ ይሓይሽ' ብምባል ክንደይ ከምእትኸፍሎም ዋጋ ዕዳጋ ከምዝኣትዉ መገሽኡ ተደጋጋሚ ጸገም ዘጋጠሞ መነስ ይሕብር። ካልኦት ሃገራት ኣፍሪቃ'ኸ?

Figure 8. Output summary System.

Reference/ Manual Summary

Refetrence: ሞት ጸሊም ብጻዕዳን ጸሊምን እንታይ እዩ ፍልልዩ?ብሕብሪ ቆርበቱ ጥራሕ ከም ገበነኛ ተቖጺሩ ብፖሊስ ብጭካነ ዝተቐትለ ጸሊም ኣመሪካዊ ጆር
ጅ ፍሎይድ ኣብ ሂዩስቴን ቴክሳስ ተቐቢሩ።ጨካን ተግባር ኣብ ኣእምሮ ኩሉ'ቲ ነቲ ፍጻሜ ብተንቀሳቓሲ ስእሊ ዝረኣየ ዘበለ ተሓቲሙ ኣሎ።ቅትለት ጆርጅ ፍ
ሎይድ፡ ኣብ ኣመሪካን ዝተፈላለየ ክፍልታት ዓለምን፡ እቲ ኣብ ልዕሊ ጸለምቲ፡ ንዘበናት ዝሰረተ ምምሕዳራዊ ስርዓት ዓሌታዊ ኣድልዎን ወጽዓን ንከብቅዕ ዝጽው
ዕ ሰፊሕ ህዝባዊ ምንቅስቓስ ኣለዓዒሉ'ዩ።ኣብ ዝተፈላለያ ሃገራት ኣፍሪቃ ብፖሊስን ብሓይልታት ጸጥታን ኣብ ልዕሊ ህዝቢ ዘጋጥም ግህሰት ብኣምኒስት ኢንተ
ርናሺናልን ሂዩማን ራይትስ ዎችን ብዙሕ ተቓውሞታት ክገጥሞ ጸኒሑን ኣሎንኣብ ኢትዮጵያ ኦሮምያን ኣምሓራን ከተማ መቐለ እውን፡ ኣብ ዝተፈላለየ እዋን፡
ትማልን ሎምን፡ ሓይልታት ጸጥታን ፖሊስን ኣብ ልዕሊ ሰለማዊ ሰባት ዝተፈላለየ በደላት ፈጺሞምን ይፍጽሙን ኣለው።ስለምንታይ እዮም ግን ሕጊ ከኽብሩ ዝግ
በኦም ፖሊስ ሕጊ ዝጥሕሱ እሞ ኸኣ ሰባት ዝቐትሉ?እቲ ገበን ፈጺሙ ተባሂሉ ዝኸሰሰ ፖሊስ ኣብ ቤት ፍርዲ ምስ ቀረበ ምስቲ ዝፈጸሞ በደልን ግህሰትን ዝዳረ
ግ ፍርዲ ስለዘይወሃቦ፡ ብሓውሲ ዕበጦ ዕበጦ ስለዝውዳእ ጸለምቲ ኣፍሪቃውያን ኣመሪካውያን ፖሊስ እናኸርእዩን ብዝኾነ ይኹን ምኽንያት ደው ክብሉ ምስ ዝሓ
ቶም ከይትኩሰሎም ከምዝጭነቑ እዮም ዝዛረቡ።ኣምኒስት ኢንተርናሺናል ተመራማሪ ምብራቕ ኣፍሪቃ ዓብዱላሂ ሃላኪ፡ ፖሊስ፡ ኣማራጺ ኣብ ዝስእኑሉን ህይወት
ብዙሓት ንምድሓን ሓይሊ ክጥቀም እንተዘይተገዲዱ፡ ኣብ ዝኾነ ዝፍጥር ዕግርግር ንህዝቢ ክከላኸለን ክረጋግእን እምበር ንህዝቢ በባትር ኮነ ብተተካሲ ክቐትል
ኣይግባእን'ዩ ይብል።ሓደ ካብቲ ጭካነ ፖሊስ ክቑጸሩ እንከለውን ምስተቓጽሩን ንመሰላት ዜጋታት ዘኽብር ብቑዕን ዓሚቕን ስልጠናን ቀጻሊ ትምህርትን ስለዘይ
ወሃቡ ኪኾኑ ይኽእሉ እዮም ዝብሉ ወገናት እውን ኣለው።መብዛሕትኦን ሃገራት ንሰራዊትን ፖሊስን ዝህበኦ ስልጠና እምበር ነዚ ሓይሊ እዚ ንኪቖጽራ ዝጥቀማሉ
ረቛሒ ዝተፈላለየ እዩ ዝብሉ ተንተንቲ እዚ ዛዕባ፡ ኣብተን ብዙሕ ዘይማዕበላ ሃገራት ሳልሳይ ዓለም ፖሊስ ነቲ ስልጣን ዝሓዘ ስርዓት እምበር ብሰልፊ እምቢትኡ
ንዝገልጽ ህዝቢ ከም ጸላኢ እዮም ዝርእዩዎ ዝብል ምጉት ዝለዓል እንተኾነ፡ ፖሊስ ኣመሪካ'ኸ እንታይ ረኺቦም ዝብል ጉዳይ ክለዓል እምበኣር ባህሪያዊ ይኸውን
።እዚ ኸኣ ኣብዛ ዓለም ገና ንዓሌታዊ ማዕርነትን ፍትሕን ዝግበር ቃልሲ ነዊሕ ግዜን ጻዕርን ከምዝሓትት መረጋገጺ እዩ ይብል ዶክተር ተስፍኣለም ነቲ ጸሊም
ኣመሪካዊ ብስርዓተ ትምህርቲ፡ ጥዕናን ፍትሕን ዝበጽሖ ኣድልዎ ከም ኣብነት ብምቕራብ።

Figure 9. Reference / Manual Summary.

3.4. Summary Evaluation Measures

For experimentation and evaluation, a number of factual reports from selected news articles with variable numbers of sentences were used. On each of those, the proposed algorithm was applied, and the ROUGE evaluation metric was used to evaluate the precession, recall and f-score of the system-generated over the reference/manual summary.

In this point, the Feature Extraction and feature Enhancement are carried out as explained in sections 3.2.1 and 3.2.2 for the given News article. For every sentence of the given document, the values of the feature vector sum and enhanced feature vector sum have been generating the final summary as displayed in Figure 8. This was extracted the hierarchical representation of the data to enhance the features, that, at first did not have much variety, and subsequently finding the latent factors using algorithm Restricted Boltzmann Machine (RBM). The sentences have at that point been ranked on the basis of the final feature vector sum and summaries are created as proposed in section 3.6.

Then after. the final summary was evaluated using ROUGE evaluation metrics. The ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially a set of metrics for evaluating automatic summarization. It works by comparing an automatically produced system summary against a set of reference summaries (typically human-produced). Let's say that we have the following system and reference summaries:

The ROUGE evaluation for the above sample summary was displayed as below Figure 10 and the finall summary evaluation conducted in for different documants/ articles was summarise in following Table 4.

```
[{
'Rouge-1':
        {'r': 0.5064935064935064,
         'p': 0.30077120822622105,
         'f': 0.37741935016342354},
'Rouge-2':
        {'r': 0.31802120141342755,
         'p': 0.18518518518518517,
         'f': 0.23407021641474504},
'Rouge-l': |
        {'r': 0.341991341991342,
         'p': 0.20308483290488433,
         'f': 0.2548387050021333}
}]
```

Figure 10. ROUGE-Score the extracted system summary over the reference summary.

Table 4. *Summary Evaluation Result.*

Document	ROUGE Evaluation Metric								
	Rouge-1			Rouge-2			Rouge-l		
	R	P	F	R	P	F	R	P	F
Doc-1	0.50	0.30	0.37	0.31	0.18	0.23	0.34	0.20	0.25
Doc-2	0.45	0.51	0.48	0.27	0.36	0.31	0.43	0.48	0.45
Doc-3	0.49	0.32	0.39	0.33	0.21	0.26	0.36	0.24	0.29
Doc-4	0.52	0.43	0.47	0.35	0.30	0.32	0.46	0.38	0.42
Average	0.49	0.39	0.42	0.32	0.26	0.28	0.39	0.33	0.35

4. Results and Discussion

Most of the related work explained in table 1 focus on the single or two statistical approaches of machine learning algorithms like term frequency, Probabilistic Latent Semantic Analysis (PLSA), term frequency, TopicLSA, sentence position, and Sentence Rank (SR). In this paper, we used combinations of machine learning approaches for the features extractions using the number of the thematic words to find the top most frequent words, sentences position, sentences length, sentences position relative to paragraph, numbers of proper nouns, number of numerals, number of named entities and the Term Frequency-Invers Sentences Frequency (TF ISF) and the deep leering algorithms for feature enhancement using Restricted Boltzmann Machin (RBM) was used to extract and generate the summarization.

The Restricted Boltzmann Machine (RBM) architecture of deep neural network rather than the statistical approach as explained in section 3.2.2, it uses the two layered structures called visible and hidden layers to learn and extract feature of the give machine learning approaches. The RBM was work with unsupervised and extractive single document summarization achieves satisfactory performance compared to the related works for the selected language.

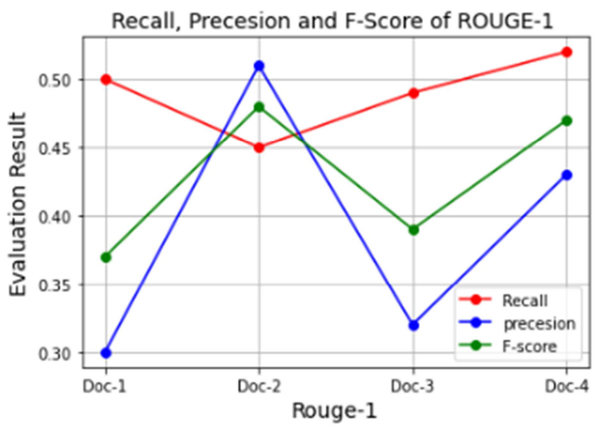

Figure 11. *ROUGE-1 Summary Evaluation corresponding to summaries of various documents.*

The result in Table 2 show that, ROUGE-1, ROUGE-2 and ROUGE-L scores for different news article corpus test set and results are averaged over the given articles.

The experiment shows that extract the most relevant information from the source news article, comparatively, the ROUGE-1 shows better average result of recall, precession and F-score test set summary. The average result of ROUG-1 shows 49% for recall, 39% precession, 42% for F-score and for the ROUGE-2 shows that 32% recall, 26% precession and 28% for F-score, and finally, for the ROUGE-l also shows that 39% of recall, 33% of Precession, and 35% of F-scores. As displayed in the following pictures shows the scores of ROUGE-1.

To answer the research question, Does the selected model was generated well organized and coherent summary? For testing and evaluation, a variety of factual reports from different news articles with variable numbers of sentences were employed. On each of those, the suggested model was applied, and the system-generated summaries were compared to the summaries created by humans.

The models for the Feature Extraction and Enhancement were conducted as anticipated in sections 3.2.1 and 3.2.2 for all the given documents. The values of feature vector sum and enhanced feature vector sum for each sentence. The Restricted Boltzmann Machine has extracted a hierarchical representation out of data that initially did not have much variation, hence discovering.

For the research question, to what extent the system summary was efficient compared to the manual summary? The coherence of the summary shows the precession or how much of the system summary was relevant or needed, and this shows different result for different documents (30%, 51%, 32% and 43% for Doc-1, Doc-2, Doc-3 and Doc-4 respectively). As the harmonic mean of the system's precision and recall values of ROUGE-1 was 42%, which shows the mean average of extracted summary with respect to the refence summary was coherent and well organized.

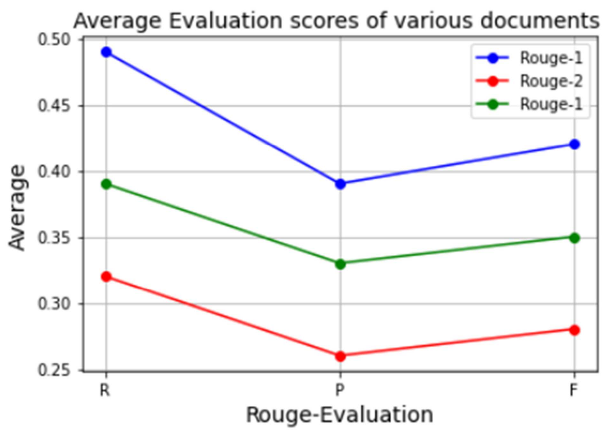

Figure 12. *Average Evaluation scores of various documents.*

For the research question, Does the extractive text summarization approach was properly identified as salient and coherence summary in the original document? This shows as the result displayed in Figure 12, the various documents have different scores. In this case, from the given documents, News Article 'Doc-2' have the higher scores in Rouge-1 of the overlaps of the system summary to the reference summary as recall 45%, Precession 52% and F-score 48%. Here, the precession has the higher score, and this shows, extractive text summarization was properly identified as salient and coherence summary with in the original document.

5. Conclusions

In this paper, the model RBM was used as an unsupervised learning algorithm for enhancing the accuracy of the summary. It was noted that the suggested method produces concise summaries of the given single news article without any irrelevant words. By the features such as Sentence-Centroid similarity and thematic words used in the feature extraction stage was used to improve the connectivity of the sentences. This also helps the proposed model to produce concise and clear summary. In this work, the proposed model scores based on the ROUGE evaluation shows an average of 49% recall, 39% precession and 42% F measure was obtained in Rouge-1, 32% recall, 26% precession and 28% F measure was obtained in Rouge-2 and 39% recall, 33% precession and 35% F measure was obtained in Rouge-l.

The results produced using the proposed method give better evaluation parameters in comparison with prevailing RBM method. This shows that, the evaluation score of the system summary compared to the refence summary gives higher result in Rouge-1 and the F-score or harmonic mean of precision and recall is 42% and it solves the problems of information overloading in the ever-increasing available news articles by generating the extractive summarizations.

References

[1] V. K. Dijana Kosmajac, "Automatic Text Summarization of News Articles in Serbian Language," 18th International Symposium INFOTEH-JAHORINA, pp. 20-22, 2019.

[2] O. S. G. B. H. H. C. N. a. C. P. J. K. Yogan, "A review on automatic text summarization approaches," Journal of Computer Science, vol. 4, no. 12, p. 178–190, 2016.

[3] A. Negash, "The Origin and Development of Tigrinya Language Publications (1886 - 1991) Volume One," University Library, vol. I, p. 131, 2016.

[4] A. S. Akshay Kulkarni, "Natural Language Processing Recipes," in Unlocking Text Data with Machine Learning and Deep Learning using Python, Bangalore, Karnataka, India, acid-free paper, 2019, p. 253.

[5] P. Anttila, "Automatic Text Summarization," UNIVERSITY OF TURKU, Department of Future Technologies, Computer Science, 2018.

[6] J.-M. Torres-Moreno, Automatic Text Summarization, Great Britain, the United States: ISTE Ltd and John Wiley & Sons, Inc, 2014.

[7] J. E. I. a. T. H. Gary Miner, "Practical text mining and statistical analysis for non-structured text data applications," Academic Press, 2012.

[8] H. P. Luhn, "The automatic creation of literature abstracts.," IBM Journal of research and development, vol. 2 (2), p. 159–165, 1958.

[9] H. P. Edmundson, "New methods in automatic extracting," Journal of the ACM J (JACM), vol. 2, no. 16, p. 264–285, 1969.

[10] E. T. V. Leo Laugier, "Extractive Document Summarization Using Convolutional Neural Networks - Reimplementation," Department of Electrical Engineering and Computer Sciences.

[11] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv, vol. 1408., no. 5882, 2014.

[12] M. J. E. P. Yong Zhang, "Extractive Document Summarization Based on Convolutional Neural Networks," IEEE, Singapore, Australia, 2016.

[13] M. J. E. a. R. Z. Y. Zhang, "Multi-document extractive summarization using window-based sentence representation," in Computational Intelligence, 2015 IEEE Symposium Series on, p. 404–410, 2015.

[14] J. O. A. M. C. M. M. P. L. H. C. &. J. C. F. B. De Sordi, "Design science research in practice: What can we learn from a longitudinal analysis of the development of published artifacts?," Informing Science: The International Journal of an Emerging Transdisciplin, no. 23, pp. 1-23, 2020.

[15] J. v. &. H. A. &. M. A. Brocke, "Introduction to Design Science Research," In book: Design Science Research, 10.1007/978-3-030-46781-4_1., pp. 1-13, 2020.

[16] M. T. &. K. K. Bharti Sharma, "Extractive text summarization using F-RBM," Journal of Statistics and Management Systems, pp. ISSN 2169-0014, 2020.

[17] S. V. a. V. Nidhi, "Extractive Summarization using Deep Learning," Delhi Technological University: arXiv: 1708.04439v2 [cs. CL] 9 Jan 2019, no. arXiv: 1708.04439v2 [cs. CL] 9 Jan 2019, 2019.

[18] a. M. S. Masoud Fatemi, "Joint Sentiment/Topic Modeling on Text Data Using Boosted Restricted Boltzmann Machine," in Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran, 2017.

[19] R. R. K. V. R. Ghosh, "A novel deep learning architecture for sentiment classification," 10.1109/RAIT.2016.7507953, 2016/03/01, pp. 511- 516.

[20] G. A. Birhanu, "Automatic Text Summarizer for Tigrinya Language," Addis Ababa Unversity, Addis Ababa, 2017.