Academic City University College – Accra Ghana
Society for Multidisciplinary & Advanced Research Techniques (SMART) Africa
Tony Blair Institute for Global Change
FAIR Forward – Artificial Intelligence for All - Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Accra Bespoke Multidisciplinary Innovations Conference (ABMIC)

& The Africa AI Stakeholders' Summit 14th December, 2021

Enhancing Learning Retention with Document Summarization Using NMF and ACO

Rufai M. M[1]., Sodiq K.A[2]. & Ige S. A.[3].
[1]Department of Computer Technology, Yaba College of Technology, Yaba, Lagos State, Nigeria
[2]Department of Computer Engineering, Yaba College of Technology, Yaba, Lagos State, Nigeria
[3]Department of Mathematics, Yaba College of Technology, Yaba, Lagos State, Nigeria
E-mail: m_rufai@yahoo.com

Proceedings Citation Format

Rufai M. M., Sodiq K.A. & Ige S. A.3. (2021): Enhancing Learning Retention with Document Summarization Using NMF and ACO. Proceedings of the Accra Bespoke Multidisciplinary Innovations Conference. University of Ghana/Academic City University College, Accra, Ghana. December 2021. Pp 195-210. www.isteams.net/ghanabespoke2021.
DOI https://doi.org/ 10.22624/AIMS/ABMIC2021-V2-P15

Enhancing Learning Retention with Document Summarization Using NMF and ACO

Rufai M. M[1]., Sodiq K.A[2]. & Ige S. A.[3].

ABSTRACT

Retaining learning is a desirable value for the learned. The traditional approach to learning retention is to read, understand and summarise. This explains the reason why vital points in voluminous book are highlighted for the purpose of retention. This study developed an unsupervised extractive summarization algorithm using Non-negative Matrix Factorisation and Ant-Colony Optimization Techniques for electronic text summarization. The developed system is an improvement over the existing NMF and LSA in the sense that LSA and NMF do not adequately address the noise issue that features in semantic document representation thereby leading to poor selection of meaningful sentence that represent the document summary. While NMF was used for factorising the initial Document-Term matrix generated from the document, ACO was used to remove the remnant of noise from the DTM. The improved algorithm is applied on a literature text material which reduces the voluminous material into handy summary by extraction. The algorithm was evaluated using compression ratio, retention ratio and was found to be adequate.

Keywords: Text Summarization, Non-Negative Matrix Factorisation, Ant-Colony Optimization

1. INTRODUCTION

Retention can be defined as the accuracy with which a reader recall the information read over a period of time(1). Summarization has been recommended as one of the strategies for retaining read or studied text. A summary captures substantial part of a document using less text. Summarised text are easy to retain during study and recall during test. Hence, the strong relationship between summary and retention. Text summary is a text that is extracted from a collection of texts, that contains a significant portion of the information in the original text and that is not longer than half of the original text(s) (Allahyari et al., 2017).

Several methods have been deployed in recent time in the summarization of documents. Notably amongst these methods is supervised and unsupervised methods. Supervised methods consist of algorithms that uses a large amount of human made summaries to train an artificial system for text summarisation. Unsupervised methods in contrast to supervised, select important sentences from a document without using labelled summaries during training (Lee, Park, Ahn & Kim 2009). The purpose of document summarisation is to produce an extract of the original text that is shorter and consumes less reading time without compromising the important semantic contents. Document summarization are classified into different types (Nenkova and McKeown, 2012; Saggion and Poibeau, 2013) based on input type, purpose and output type. Input type describe whether the input document is a single document or multiple documents. Classification according to purpose has 3 classes which are generic, domain specific, or query-based.

Generic is summarization with no regard for topics or domain, while domain specific summaries have recourse to the domain in which the summary is based on e.g. summarization of finance articles. Query-based summary contains only users specified information. Two basic types of text summaries based on output type exist which are extractive and abstractive summaries. Extractive summaries extract a section of the original documents that best represent the original document semantically. This extracts usually include important sentence from the original document. (Aliguliyev, 2009; Ko and Seo, 2008). Abstractive summarization, the selects from the original documents which are later combined in a compressed form such that unimportant sections of the sentences are eliminated. (Ganesan et al., 2010; Khan et al., 2015).

This study will focus on extractive, multi-document, domain specific document summarization and primarily develop an algorithm that will improve term extraction, document semantic representation and dimension reduction. In this paper, we propose a new algorithm for multi-document, domain specific document summarization method using Non-negative Matrix Factorization and Ant-Colony Optimization techniques. The proposed method inherits the benefits of unsupervised method by having no need for training summaries. Secondly, the term-sentence vectors are treated for noise using the Ant Colony optimization techniques which further improve the quality of the summary output.

The remainder of this paper is organized as follows: Section 2 describes related work regarding document summarization; Section 3 describes NMF and ACO; Section 4 describes NMF-ACO summarization approach; Section 5 describes the performance evaluation. Finally, in Section 6, we conclude the paper with directions for future research.

2. RELATED WORKS

Radev et al. (2004) presented the MEAD algorithm as a platform for multidocument text summarization. The MEAD is a public domain use for multidocument, multilingual summarization. It has been widely used by more than 500 companies and organizations. It has been applied in various applications ranging from the mobile phone technologies to web page for summarization and also for novelty identification.

Ju-Hong, Sun, Chan-Min and Daeho (2009) applied an unsupervised method for automatic generic document summarization using Non-negative Matrix Factorization (NMF) to select sentences for the document summarization. No training summaries were required and the semantic feature vectors extracted were interpretable. More meaningful sentences for generic document summarization were selected using the NMF method than those selected using LSA. Baruque and Corchado (2010) presented a weighted voting summation of Self-Organization Maps (SOMs) ensembles. Weighted voting superposition was used for the outcomes of an ensemble of SOM. The objective of this algorithm was to attain the minimal topographic error in the map. Hence, a weighted voting process was done among the neurons to calculate the properties of the neurons located in the resulting map.

Alguliev et al. (2013) applied an evolutionary optimization algorithm to reduce redundancy in multidocument summarization. The algorithm creates a summary by collecting the salient sentence from the multiple documents. This approach uses the summary to-document collection, sentence-to-document collection, and sentence-to-sentence collection to choose the most important sentences from the multiple documents.

Hence, it reduces the redundancy in the document summary. According to the individual fitness value, the algorithm can adaptively alter the crossover rate.

R. M. Aliguliyev, and N. R. Isazade (2013) also proposed constraint driven document summarization models. This approach was developed based on the following two constraints: (1) diversity in summarization to reduce the redundancy in the summary among the sentences; (2) sufficient coverage to avoid the loss of document major information to generate the summary. In order to solve the Quadratic Integer Programming (QIP) problem, this approach utilized a discrete Particle Swarm Optimization (PSO) algorithm.

Xiong and Lu (2014) introduced an approach for multidocument summarization using Latent Semantic Analysis (LSA). Among the existing multidocument summarization approaches, the LSA was a unique concept, which uses the latent semantic information rather than the original features. It has chosen the sentence individually to remove the redundant sentences.

Su and Xiaojun (2014) discussed an extractive multidocument summarization. This approach uses the semantic role information to improve the graph based ranking algorithm for summarization. The sentence was parsed to obtain the semantic roles. The SRRank algorithm was proposed to rank the sentence, words, and semantic role simultaneously in a heterogeneous ranking manner.

Ma and Wu (2014) modelled a multidocument summarization technique using a combination of n-gram and dependency word pair. The dependency word pair identifies the syntactic relationships among the words. Each feature reproduces the cooccurrence of the common topics of multidocuments in diverse perspective. The sentence score was estimated based on the weighted sum of the features. Finally, the summary was extracted by retrieving the salient sentences based on the higher significance score model.

Nedunchelian Ramanujam and Manivannan Kaliappan (2016) presented a new concept to document summarisation using timestamp approach combined with Naïve Bayesian Classification approach for multidocument text summarization. The timestamp was used to ordered the look of the summary thereby making it coherent. More relevant information was extracted and higher linguistic qualities such as readability and comprehensibility were estimated using scoring strategy inherent in the approach. The approach was compared with the MEAD algorithm and was observed to be faster in generating results. Moreover, the proposed method results in better precision, recall, and F-score than the existing clustering with lexical chaining approach.

2.1 Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF) is a low rank approximation technique with reduced storage, run-time requirements, reduced redundancy and noise. It allows for additive parts-based, interpretable representation of the data. NMF approximates a matrix V by

$$V_{n\times m} \approx W_{n\times r} H_{r\times m} \tag{3.10}$$

where W and H are NMF factors and all entries in V, W and H are to be non-negative.

r,m,n represent the rank of the matrix r which is chosen to satisfy $(n+m)r < nm$

The goal of NMF is to minimize the original matrix V. The objective function used is the Frobenious Norm shown in equation 3.11

$$min\|V - WH\|_F^2 = min \sum\sum\left(V_{ij} - (WH)_{ij}\right)^2 \quad (3.11)$$

NMF perfectly fits in as a better alternative to SVD in dimension reduction in LSA because of its sparsity and non-negativity; reduction in storage and its interpretability. However, its major challenge is its convergence issue because different NMF algorithm can converge to different local minima. This challenge is addressed by choosing the right initialization and update strategy. In this research the problem of dimension reduction is tackled by using two strategies: the first strategy is to use SVD-LSA to initialize the factors for minimizing NMF objective function prior to factorization while the second strategy seeks to iteratively improve the quality of the dimension reduction accuracy using Ant Colony Optimization Technique.

2.2 Ant Colony Optimization (Aco)

Ant Colony Optimization (ACO): is a swarm intelligence algorithm that leverage on the foraging behaviour of ant in their search for food in finding optimal path. The purpose of the ACO optimization strategy is to determine an optimal value of W and H for which the distance matrix is minimized. The Frobenius norm is the objective function. The choice of the Frobenius norm as the objective function is based on the fact that it offers some properties that are beneficial for combining NMF and optimization algorithms. There exist a matrix D which refers to a distance matrix storing the distance between the original data and the approximation (WH), $D = Z - WH$. The Frobenius norm of a matrix $D \in \mathbb{R}^{m \times n}$ is defined as:

$$\|D\|_F = \left(\sum_{i=1}^{\min(m,n)} \sigma_i\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}^2\right)^{1/2} \quad (3.22)$$

The row-wise computation of the Frobenius norm is given as:

$$\|D\|_F^{RW} = \left(\sum_{i=1}^{m} |d_i^r|^2\right)^{1/2} \quad (3.23)$$

where $|d_i^r|$ is the norm of the ith row vector of D and r_j^i is the jth element in row i

while the column wise calculation is

$$\|D\|_F^{CW} = \left(\sum_{i=1}^{m} \left|d_j^c\right|^2\right)^{1/2} \quad (3.21)$$

where $\left|d_j^c\right|$ is the norm of jth column vector of D

The ACO techniques iteratively improves the quality of NMF factors that gives a better approximation of the weighted Document-Term Matrix. It does this by iteratively searching for the factors of NMF i.e. W and H which satisfy the condition of a minimal distance D, which is a measure of the error between Z and WH. The optimization is performed on selected rows/columns of W/H. For the effectiveness of the approximation and its fast convergence, the rows of D with the highest norm are identified and the corresponding rows of W and columns of H that minimize the Frobenious Norm objective function are sought for using ACO. Algorithm 3.2 shows the steps involved in each of the phases.

3. METHODS

An algorithm which integrates Non-Negative Matrix Factorisation with Ant Colony Optimization technique was developed to perform multidocument summary on the collected news material. The improved algorithm collects text material on a topic, converts these documents into a document-term matrix, performs weighting of the matrix using term frequency-inverse document frequency, and factorise and initialise using Single Value Decomposition(SVD) and Non-Negative factorization technique. The resultant matrix is further treated for noise by applying Ant Colony Optimisation to the weighted and factorised matrix.

The generic relevance of each component document was computed and the best document is selected as the summarised document domain of relevance. The algorithm was evaluated using compression ratio, retention ratio and a comparative performance with similar algorithms was done to prove its adequacy.

3.1 Algorithm to extract sentences from a document using NMF

24. News report on a particular event from various media were collected event were collected
25. Decompose the document into individual sentences, and let r be the number of sentences for generic document summarization.
26. Perform stopwords removal and word stemming operations.
27. Construct the term-by-sentence matrix V.
28. Perform NMF on matrix V to obtain H.
29. Optimise W and H in NMF using ACO
30. For each sentence, calculate its generic relevance.
31. Select r sentences with the highest generic relevance values as summary.

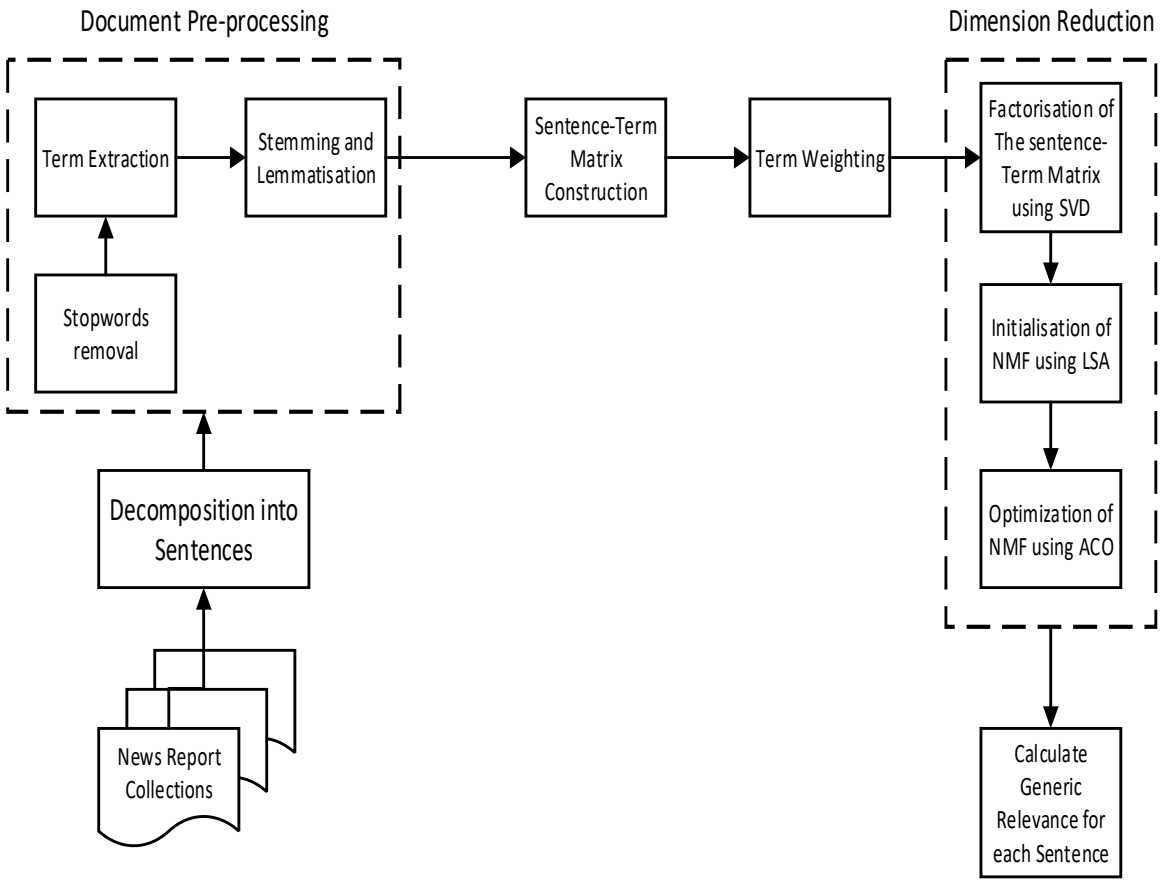

Figure 1: Block Diagram of The News Summary Algorithm

3.2 Dataset Collection

The source of the dataset are online news report on the sacking of the minister of Agriculture from various newspaper media which include The Punch, The Nation, Vanguard, PM News, Premium Times and Daily Trust. The news reports are broken down into sentences as the basic unit.

3.3 Document Pre-Processing

The next stage after document collection is Document Pre-Processing. This involves three stages which are:

i) Stop words Removal
ii) Term Extraction
iii) Stemming and lemmatization

i) Stop Words Removal

Stop words are eliminated with the objective of removing words with very low discrimination values for similarity purpose. The dictionary-based approach to stop word removal was used. A generic list of stop words was retrieved from an existing online dictionary, "rank.nl" and a MATLAB program was applied to search for each stop word in each of the collected documents and if found, removed from the document.

ii) Term Extraction

The purpose of term extraction is to generate list of terms that are relevant to the input domain. A well generated or extracted term will facilitate adequate semantic representation of the individual documents. These terms are extracted from the document collections. Two steps were applied in the term extraction. The first step ranks terms from the domain of input document to determine the most relevant terms. The second step uses ACO to train some selected score functions in order to determine the correct combination of feature weights. The procedure used in extracting relevant terms is outlined as follows:

1. Read the input document.

2. Perform the following pre-processing steps which consist of
   a. Analyse the syntactic content of every input sentence using Syntactic parser and output a list of syntactic information e.g. Noun Phrase
   b. Filter out stop words from each of the list of Noun Phrase
   c. Stem the Noun Phrase to produce list of clean Noun Phrase which are called the term candidate.

3. Associate the term candidate to a vector that contains five features.

   Use the five features to calculate the term score and then rank the terms based on their score using Equation (3.1)

$$Score(t) = \sum_{i=1}^{5} w_i . Scoref_i(t) \tag{3.1}$$

where W_i indicates the weight of f_i.

The features that were used to compute the term score and further rank the terms are:

a. **Domain Relevance (DR):** -domain relevance is described as the amount of information captured in the target document with respect to contrastive documents. D_k is the domain of interest (consisting of a set of relevant documents) and $D_1 \ldots D_n$ is sets of documents in another domain, domain relevance of a term t in class D_i is computed using Equation (3.2)

b.

$$DR(t, D_k) = \max_{i<j<n} \frac{P(t|D_k)}{P(t|D_j)} \tag{3.2}$$

Where

$$P(t|D_k) = \frac{f_{t,k}}{\sum_{t' \in D_k} f_{t',k}} \tag{3.3}$$

$f_{t,k}$ is the frequency of term t in the domain D_k
$f_{t',k}$ is the frequency of terms t' in the domain D_k

c. **Domain Consensus (DC):** measures the distributed use of a term in a domain D_k. It can be computed using Equation (3.4)

$$DC(t, D_k) = \sum_{d_j \in D_k} \left(P(t|D_k). log_2 \left(\frac{1}{P(t|D_i)} \right) \right) \quad (3.4)$$

$$P(t|D_j) = \frac{f_{t,j}}{\sum_{d_j \in D_k} f_{t,j}} \quad (3.5)$$

d. **Term Cohesion**: Cohesion is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning. Term cohesion is used to calculate the cohesion of the multi-word terms. It is proportional to the co-occurrence frequency and length of the term. It is expressed mathematically as:

$$TC(t) = \frac{|t|.log_{10}(f(t)).f(t)}{\sum_{w_i \in t} f(w_i)} \quad (3.6)$$

where w_i are the words composing the term

e. **First Occurrence (FO):** It measures the distance of a phrase from the beginning of a document in words i.e. the number of words that precede the first occurrence of the phrase from the beginning of the document.

$$FO(t, d) = \frac{\sum_{(w_i \in d: i \leq w \leq t)} f(w_i)}{\sum_{(w_i \in d)} f(w_i)} \quad (3.7)$$

f. **Length of Noun Phrase(LNP):-** It is calculated as frequency multiplied by length (in words)

$$LNP(t) = f(t). |t| \quad (3.8)$$

iii) Stemming and Lemmatization

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Stemming was done by applying the Porters Stemming rules. These terms and their synonyms are given the same weighting during term weighting generation.

3.2.3 Term-Sentence Matrix Construction

After document collection, the next stage is to construct the Sentence-term matrix. In this matrix, each term is a column and each Sentence is a row. Each cell contains the number of times that a term appears in the sentence.

3.2.4 Term Weighting Generation

The essence of term weighting is to ensure that rare words are weighted more heavily than common words. For example, a word that occurs in only 5% of the documents should probably be weighted more heavily than a word that occurs in 90% of the documents. The reason for this is because rare words reveal better similarity features among documents. The term weighting approach adopted in this research is the TF-IDF (Term Frequency - Inverse Document Frequency). Under this method, the count in each cell is replaced by the output of Equation 3.9

$$w_{t,d} = \left(1 + log tf_{t,d}\right). log \frac{N}{df_t} \quad (3.9)$$

where:

$w_{t,d}$=weight of terms and documents
$tf_{t,d}$= the frequency of term t in document d
N=Number of documents
df_t= number of documents with term t

In other words, $w_{t,d}$ assigns to term t a weight in document d that is:

1. highest when t occurs many times within a small number of documents thus lending high discriminating power to those documents when similarity between documents is observed;
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. Lowest when the term occurs in virtually all documents.

Dimension Reduction and Optimization using LSA-NMF-ACO

Dimension reduction using LSA has been observed in literatures to have the following setback:

1. It does not lead to proper approximation for the original matrix,
2. the presence of negative value in the cell of term-document matrix makes it un-interpretable.
3. It does not adequately capture the document semantic content

This research addressed these problems by improving the existing LSA through its Dimension Reduction function to further minimize noise in its Sentence-Term Matrix by introducing initialization and optimization of its factors using Non-Negative Matrix Factorization (NMF) and Ant Colony Optimization (ACO). An improved algorithm, LSA-NMF-ACO is developed by factorizing the weighted Sentence Term Matrix using SVD-LSA into three LSA factors U, S and V^T. where U and V are unitary matrix and S is a singular diagonal matrix.

The U, S, V factors of LSA were used to initialize the W and H factors of NMF. ACO optimizes these factors by searching for the value of the W and H factors of NMF that guarantees optimal reduction using the Frobenious norm of the distant matrix as the objective function. The Distant Matrix is the difference between the initial document term matrix and the reduced matrix using the improved algorithm

Generic Relevance Computation

A novel method to select sentences based on NMF and define the generic relevance of a sentence (GRS) as follows:

$$\text{Generic relevance of a jth sentence} = \sum_{i=1}^{r} \left(H_{ij}.\ weight\ (H_{i*}) \right)$$

$$\text{Weight}\ (H_{i*}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{r} \sum_{q=1}^{n} H_{pq}} \qquad (3.10)$$

Table 3 show the generic relevance of each sentence. The first four sentences with highest generic relevance are chosen as the News summary

4. RESULTS AND DISCUSSION

Term-Sentence Matrix

It is a matrix showing the number of times a particular term appears in a particular sentence as shown in Table 1

	assembl	assist	attempt	author	believ	bellc	born	british	broader	call	caus	centur	chao	chose	circl	civil	claim	clear	coalit	colleg	commun	conceiv	concern	confer	congress	consic
S1	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S2	0.00	0.00	0.00	0.00	0.00	0.08	0.27	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S3	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00
S4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S6	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00
S7	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00
S8	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.25	0.00
S10	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S11	0.42	0.00	0.00	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S13	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.00
S15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S16	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.28
S20	0.00	0.00	0.00	0.00	0.29	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S21	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S24	0.00	0.00	0.52	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	constitut	council	coup	cultur	death	democrat	deputi	descend	desir	despit	destin	develop	discuss	dispur	distast	district	dom	drew	drive	educ	elect	elev	emir	end	enter	etat
S1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00
S4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.32	0.00
S5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S6	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00
S7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00
S8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00
S9	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00
S12	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S13	0.00	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00
S14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S15	0.00	0.00	0.00	0.00	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S19	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S20	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S21	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00
S23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00
S24	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26

The table shows 183 terms are extracted from 24 sentences collections that make up the life history of Sadauna of Sokoto

The entries in the table are subjected to weighting using equation 3.9 and this generates the result shown in table 1. The essence of weighting is to give prominence to terms that contribute more to the semantic meaning of the sentence. Terms with higher weight have higher discrimination value hence determine the semantics of the sentence in which they appear.

Table 2 shows sentences and there computed generic relevance using 3.10. The generic relevance shows the extent by which each sentence represent the document summary. The first four sentences having the highest generic relevance value (i.e. S6,S12, S24, S18,S8) are extracted as the document summary.

1https://www.ctsnet.edu/the-formula-for-successful-learning-retention/

Table 2: Sentences Generic Relevance

SN	SENTENCE	GENERIC RELEVANCE
S9	In 1948 this organization affiliated with the newly founded Northern People's Congress (NPC), originally conceived as a cultural organization but destined to become the leading political party in Northern Nigeria.	0.006156
S8	In 1945 he assisted in the formation of the Youth Social Circle in Sokoto, a discussion group of Northern educators and civil servants.	0.006271
S7	As World War II drew to an end, Bello became involved in broader political concerns.	0.005011
S6	The new sultan immediately conferred upon him the traditional, now honorary, title of sardauna and elevated him to the Sokoto Native Authority Council.	0.009265
S5	In 1938 he made an unsuccessful claim to the office of sultan of Sokoto.	0.003208
S4	In 1934, after teaching several years in the Sokoto Middle School, he entered the emirate administration as district head of Rabah.	0.003961
S3	Bello received his education first at the Sokoto Provincial School, then at Katsina Teacher Training College.	0.006052
S24	Bello's attempt to support his political allies on this occasion was the immediate, though not sole, cause for an attempted coup d'etat in January 1966, during which Bello was assassinated.	0.007547
S23	In the fall of 1965 the NNDP claimed victory in a hotly disputed regional election, and the Western Region lapsed into chaos.	0.00331
S22	The coalition party, called the Nigerian National Alliance, won a clear majority in the federal elections of 1964.	0.006081
S21	In 1964 Bello led the NPC into an alliance with the Nigerian National Democratic Party (NNDP) of the Western Region.	0.005853
S20	Therefore, when Nigeria became independent in 1960, Bello chose to remain premier of the Northern Region, while the deputy president of the NPC, Alhaji Sir Abubakar Tafawa Balewa, became prime minister of the Federation.	0.003208
S2	Ahmadu Bello was born in Rabah, North West State, a descendant of Uthman don Fodio, the renowned 19th-century Moslem leader of Northern Nigeria.	0.004053
S19	Although he participated in national discussions on constitutional reform and from 1952 to 1959 was a member of the Federal House of Representatives, he was concerned primarily with the development of the North and the protection of that region from what he considered Southern incursions.	0.004053
S18	He had an expressed distaste for the Southern style of politics and had no desire for participation in the federal government, which would require his residence in Lagos.	0.006551

SN	SENTENCE	GENERIC RELEVANCE
S17	Despite this, his role in national politics remained anomalous.	0.004943
S16	As president of the NPC and premier of the Northern Region, Bello was perhaps the most politically powerful person in Nigeria during the first 5 years of independence.	0.003719
S15	In 1954 he became the first premier of Northern Nigeria, a position he held until his death.	0.003427
S14	In the following year he accepted the regional portfolio of community development and local government.	0.005763
S13	In 1952 in the first elections held in Northern Nigeria, he was elected to the Northern House of Assembly, where he became a member of the regional executive council and minister of works.	0.004071
S12	During the 1949-1950 discussions of constitutional reform he became a leading spokesperson for the Northern view of federal government.	0.007547
S11	In 1949 he was elected by the Sokoto Native Authority to the Northern House of Assembly.	0.003719
S10	Bello became increasingly active in the NPC and ultimately its president.	0.003517
S1	The Nigerian political leader Alhaji Sir Ahmadu Bello (1909-1966) was the leading Northern spokesman during Nigeria's drive to gain independence from the British.	0.004059

The text below are the extracted document summary:

Table 3: The Extracted summary based on Generic Relevance

SN	SENTENCE	GENERIC RELEVANCE
S6	The new sultan immediately conferred upon him the traditional, now honorary, title of Sardauna and elevated him to the Sokoto Native Authority Council.	0.009265
S12	In 1945 he assisted in the formation of the Youth Social Circle in Sokoto, a discussion group of Northern educators and civil servants.	0.007547
S24	He had an expressed distaste for the Southern style of politics and had no desire for participation in the federal government, which would require his residence in Lagos.	0.007547
S18	During the 1949-1950 discussions of constitutional reform he became a leading spokesperson for the Northern view of federal government.	0.006551
S8	Bello's attempt to support his political allies on this occasion was the immediate, though not sole, cause for an attempted coup d'etat in January 1966, during which Bello was assassinated.	0.006271

The new sultan immediately conferred upon him the traditional, now honorary, title of Sardauna and elevated him to the Sokoto Native Authority Council. In 1945 he assisted in the formation of the Youth Social Circle in Sokoto, a discussion group of Northern educators and civil servants. He had an expressed distaste for the Southern style of politics and had no desire for participation in the federal government, which would require his residence in Lagos. During the 1949-1950 discussions of constitutional reform he became a leading spokesperson for the Northern view of federal government. Bello's attempt to support his political allies on this occasion was the immediate, though not sole, cause for an attempted coup d'etat in January 1966, during which Bello was assassinated.

5. EVALUATION

Two performance metrics were used in evaluating the algorithm which were Retention Ratio and Reduction Ratio

The Retention ratio (R) measures the quantity of the original text preserved in the summary. It is computed by calculating the similarity value between the individual text and the generated summary text. The average of this similarity value gives the retention ratio AvgR.

$$R=cosSIM\left(\overrightarrow{S_1}.\overrightarrow{S_2}\right)=cos(\theta)=\frac{\overrightarrow{S_1}.\overrightarrow{S_2^n}}{\left\|\overrightarrow{S_1}\right\|\left\|\overrightarrow{S_2^n}\right\|}$$

$$AvgR=\frac{\sum_1^n coSIM}{n}$$

The calculated average retention ration =0.83 which implies that 83% of the entire text is captured by the summary which is a significant value

The Reduction ration is RR

$$RR=\frac{number\ of\ words\ in\ summary}{number\ of\ words\ in\ original\ document}=\frac{123}{533}=0.23$$

6. CONCLUSION

The algorithm shows a better performance on multiple document summarization with reference to the evaluation criteria i.e. Reduction ratio and retention ratio hence it can be used for summary related applications.

REFERENCES

1. Lee, J. H., Park, S., Ahn, C. M., & Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, *45*(1), 20-34.
2. Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence
3. rewriting. In ACL (1), pages 675–686. Association for Computational Linguistics.
4. Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive
5. recurrent neural networks. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016.
6. Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In EMNLP, pages 4098–4109. Association for Computational Linguistics
7. Kuppan, S., & Sobha, L. (2009, June). An Approach to Text Summarization. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)* (pp. 53-60).
8. Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 259–264.
9. Mehdi Allahyari and Krys Kochut. 2016. Discovering Coherent Topics with Entity TopicModels. InWeb Intelligence (WI), 2016 IEEE/WIC/ACMInternational Conference on. IEEE, 26–33.
10. MehdiAllahyari and Krys Kochut. 2016. Semantic Context-Aware Recommendation via TopicModels Leveraging Linked Open Data. In International Conference on Web Information Systems Engineering. Springer, 263–277.
11. Mehdi Allahyari and Krys Kochut. 2016. Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network. In Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on. IEEE, 63–70.
12. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. ArXiv e-prints (2017). arXiv:1707.02919
13. Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. Expert Systems with Applications 40, 17 (2013), 6976–6984.
14. Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 481–490.
15. Horacio Saggion and Thierry Poibeau. 2013. Automatic text summarization: Past, present and future. In Multi-source, Multilingual Information Extraction and Summarization. Springer, 3–21.
16. Yogesh Sankarasubramaniam, Krishnan Ramanathan, and Subhankar Ghosh. 2014. Text summarization using Wikipedia. Information Processing & Management 50, 3 (2014), 443–461.