University of Ibadan
Journal of Science and Logics in ICT Research

Towards An Automatic Text Analysis and Summarization In Yoruba Language Using Transfer Learning Approach In Natural Language Processing

Oguntunde Toyin , **Abdulazeez Tijani**
toyin.oguntunde@gmail.com, **tijaniazeez92@gmail.com**

Department Computer Science, University of Ibadan, Ibadan, Nigeria.

Abstract
Text Summarization serves is a tool which helps the user to efficiently find useful information from immense amount of information. This tool is increasingly used in both the public and private sector such as telecommunication industry, research institutes and in web-based information retrievals. Yoruba Language, being one of the three major languages spoken in the South-Western part of Nigeria and some communities in other countries like Brazil, Cuba, Haiti, Togo, Benin Republic, Trinidad and Tobago, has been classified as a language in serious danger of extinction by UNESCO Red Book on endangered languages. There are not many researchworks in the field of natural language processing for Yoruba Language, not even any on text summarization, as far as it is known to this study.Some other times, when Internet searches are made on Yoruba subjects, the response obtained is loads of information, which is difficult for individuals to patiently read to comprehension. Therefore, this study aimed at developing a system that automatically retrieves, categorize and summarize Yoruba document as per users' need. The design of the summarization system shall be divided into 3 stages vis-à-vis: pre-processing, feature extraction and summary generation stages. The developed system will read a Yoruba document which will be broken into several paragraphs using a Paragraph Segmentation module. In the Tokenization module, paragraphs will be broken into sentences in which punctuations, special characters, and digits will be eliminated in the Normalization module and finally the sentences will be broken into words. The Stop Word Filtering module will remove the stop words and reduce the text to more useful words. The Yoruba Morphology Lexical Analyser module will process every sentence to a Subject-Verb-Object pattern. Every word in the sentence will take a tag, representing its Part of Speech (PoS) position, which will be done by the Part of Speech Tagging module. The feature extraction processes will commence by using the Keyword Frequency which checks for the relevance of each words in the document by counting how many times it occurred in the document. The keyword with the highest frequency is likely to be present in the generated summary. This study proposed an extractive automatic text summarization tool for Yoruba language using Transfer Learning in Natural Language Processing. At the end of the study, it is expected that a Summarisation System would have been developed, which could be employed in generating a concise summary of any given Yoruba text. This paper also presented views on recent techniques and approaches on automatic text summarization with focus on English, Chinese, Persian, Arabic, Spanish and Hausa texts. Finally, it discussed challenges and methodologies of Automatic Text Summarisation.

***Keywords*:** Extractive automatic text summarization, Transfer learning, Natural language processing, Machine learning, Yoruba language.

1.0 INTRODUCTION

The advent of World Wide Web has brought a vast amount of on-line information so, the number of electronic documents and textual data has become huge. Due to this fact, whenever a subject matter is being sought on the Internet, the response obtained is lots of different web pages with lots and lots of information. Therefore, this research is aimed at developing a system that programmatically recover, classify and produce short description of passages according to the users; desire. The application of NLP systems is utilized in separation of text, text mining, spelling correction, speech synthesis, speech

Oguntunde Toyin and Abdulazeez Tijani (2023).Towards An Automatic Text Analysis and Summarization In Yoruba Language Using Transfer Learning Approach In Natural Language Processing, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 10 No. 2, pp. 119 – 135.

recognition,machine interpretation and Voice Agents like Google Assistant, Alexa and Siri.

Producing the short description of passages is summarising a target text into a brief copy while still retaining its constituent information and original context (Gupta & Lehal, 2010).Text short description software helps to effectively isolate meaningful details from stacks of informational resources. This tool is increasingly being used in both the public and private sector such as telecommunication industry, research institutes and for web-based information retrievals. Text summarization is an important step for information management tasks. It solves the problem of difficulty in selecting the most important portions of the text for summarisation.

Currently, Yoruba text is being summarised by isolating the main themes of the text to be summarized manually using black ink and paper. The difficult words in the sentences are identified and then represented in a simpler form. Sometimes, proverbs in Yoruba can be used for the summary of a length text. The just described approach is tough, tedious and as far as it is known now, there has not been any automatic Yoruba text summarisation tool hence, this study is aimed at developing software that programmatically recovers, classify and produce brief form of passages on Yoruba text files depending on the desire of the user.

2.0 Approaches to Text Summarization

The goal of text summarization algorithms is to transform lengthy documents into brief and meaningful versions, which could be difficult and costly to undertake if done manually. Zhang (2008) classified methods of producing brief form of passage into four sub-division: Statistical-based, Linguistic-based, Machine learning-based and Hybrid systems as in Figure 2.1.

2.1 Statistical-Based Approaches

The statistical method allots weight to the statement depending on various yardsticks. This approach uses word frequency, uppercase words, sentence length, keywords, sentence position, and phrase structure.

Term frequency: This shows the appearance frequency of an expression in a text file. So, the statement that has highest number of repeated expression are allotted high scores. Inverse text file number of appearance produce uncommon expression appearance by computing the log reading of Term-Frequency. Term frequency, tf(t, d), is the frequency of term t,

$$\mathrm{tf}(t,d) = \frac{f_{(t,d)}}{\sum_{t \in d} f_{(t,d)}} \ldots \quad (2.1)$$

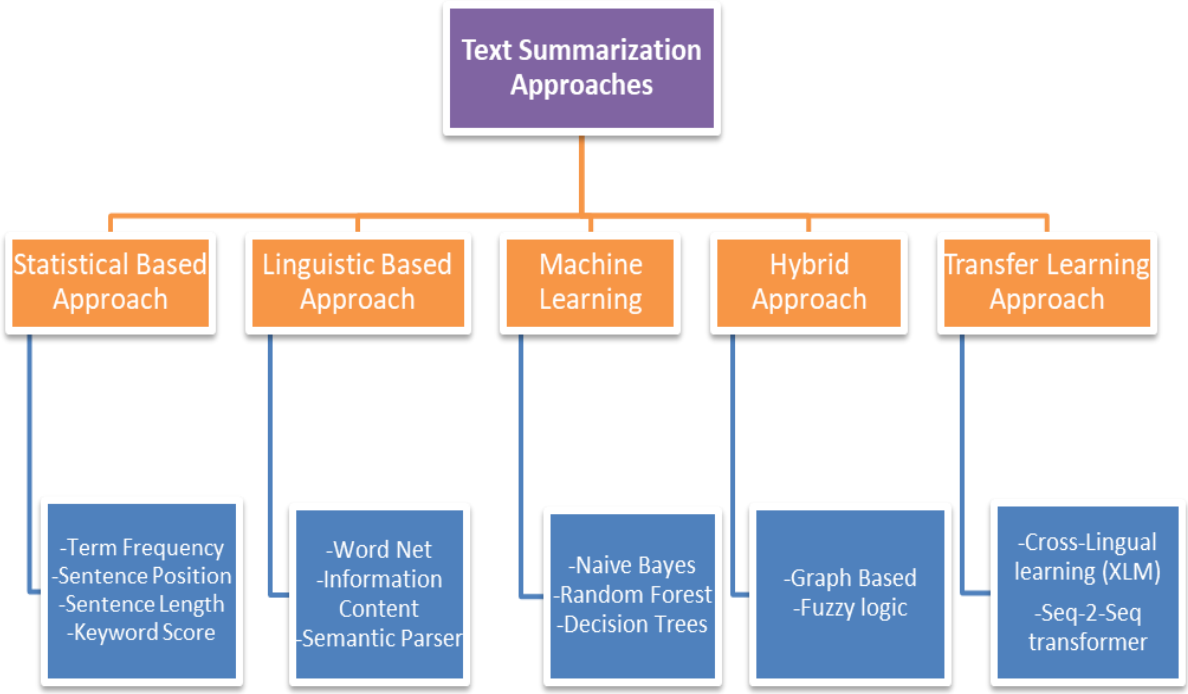

Figure 2.1: Classification of Text Summarization approaches

where $f_{(t,d)}$ stands for the number of occurrence of a word in a text file, i.e., appearance count of term, t, in a document *d*. Hence, the total count, tf(t,d) = $f_{(t,d)}$

Sentence position: This is employed to compute the weight of statement which is called position score. Position score is the location of a statement in the text file divided by the total number of sentences (Perumal *et al.*, 2011).It is mostly presumed that, significant statements lie at the start and end of a passage, wherein, the redundant onesare within the paragraph.

Sentence length: is another statistical approach that filters short sentences from the documents that might not be appropriate for a summary. This is derived from the probability of number of words in the sentence based on the length of the longest sentence. Therefore, the probability of the sentence length is determined as follows:

$$P(S, Length) = \frac{\text{number of words in the sentence}}{\text{number of words in the longest sentence}} \quad \ldots \quad (2.2)$$

whereP(S, Length) is the probability of the sentence S based on its Length.

Keyword score: employs the Term-frequency inverse sentence Frequency (TF-ISF) to assign values to the main words by computing its probability score of the frequency of the unique main words depending on the sentence length. The keywords having the greatest number of keywords counts are picked to be candidate in the short text description. Therefore, the probability of the keyword score is determined as follows:

$$P(S, keywords) = \frac{\text{number of single keywords in the sentence}}{\text{total number of words in the sentence}} \quad (2.3)$$

where P(S, keywords) is the probability score for thesentence given its keywords.

2.2 Linguistic-based Approach

Linguistic methods face challenges in employing superior linguistics tools (a discourse parser, e.t.c) and linguistic materials (Word Net, Lexical Chain, Context Vector Space, etc.). Barzilay and Elhadad (1997) projected and produced strong concept with the assistance of linguistic attributes, which a lot of memory to persist the linguistic data like WordNet and CPU capability as a result of linguistic information and complicated linguistic handling. Linguistic methods rest on the relationship, the expressions and isolating the strong ideas through analysing the expressions. Abstractive text short description depends on linguistic procedure that comprises of the semantic handling for the text short description.

WordNet: is an online lexical repository database which doubles as both a thesaurus as well as an online lookup employed by various systems by isolating correlation between expressions in English Language. WordNet are used for building both lexical chains and semantic relations. WordNet is the lexical database, specifically designed for natural language processing. Words can be categorized in WordNet to the following lexical categories such as nouns, verbs, adjectives, and adverbs however disregard prepositions, delimiters and phrases. Expressions from similar vocabulary classification, which are closely same meaning are clustered into synsets.

Information Content: This can be calculated in respect of the number of times a notion is encountered in a language resource containing text. According to Pedersen (2010), the number of appearance of a notion is updated on WordNet every time a notion is encountered as the ancestor concept number in the WordNet's order increases for the fact that each appearance unique notion indicates the appearance of common ancestor notion. Information Content is defined as the negative log of the probability of that concept (based on the observed frequency counts):

$$IC(c) = -\log P(c) \quad \ldots \quad (2.4)$$

Where c denotes the likelihood of a particular notion in WordNet. Information Content can only be computed and applied to pairs of nouns or verbs in WordNet, since these are the only parts of speech where concepts are organized in hierarchies. (Pedersen, 2010).

Semantic parsing: This is the process of translating natural language intoalogical form which is the formal meaning that is understandable by the machine.Semantic parsing is simply the task of extracting the precise meaning of a natural language. There are several

applications of semantic parsing including computer vision, automated reasoning, machine translation, question answering, ontology induction, and code generation.

2.3Machine Learning

Machine Learning can simply be defined as the process of discovering patterns in data and drawing insight from those data. Machine Learning is also the process of exploring and analysing large quantities of data to discover valid, useful, and understandable patterns in data. Classification Machine Learning techniques and algorithms such as Naïve Bayes, Random Forest, Decision Trees, Logistic Regression, Hidden Markov model are used to extract summaries of text from a given sample document. Features selection took a strategic part in developing an effective model, wherein, many features are selected to discover the relevance of a statement in text short description.

Naive Bayes: is a classification Machine learning method that is majorly employed in text classification and text short description. It is a probabilistic classifier that is based on the principle of Bayes theorem. Bayes theorem is employed to find the prospect of a hypothesis with previous knowledge. The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ldots \quad (2.5)$$

Where,
P(A|B) is the Posterior Probability of event A occurring given that B is true.
P(B|A) is Likelihood Probability of event B occurring given that A is true.
P(A) is Prior Probability of hypothesis before observing the event A.
P(B) is Marginal Probability of observing the event B.

Random Forest: is a classifier that can be used for both Classification and Regression problems in machine learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers and takes the average to improve the predictive accuracy and the performance of the model. The higher the number of trees in the forest, the better the accuracy and prevents the problem of overfitting. The pseudocode of random forest classifier is explained below:

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Decision Tree: is also a classifier that can be used for both Classification and Regression problems in machine learning, but mostly preferred for solving Classification challenges. The classifier had a tree-structure having a root node, expands to further branches and constructs a tree-like structure. The inner nodes formed the properties of the dataset, the sub-divisions gave the decision rule while the individual apex node stands for the outcome. It produces a graphical illustration to the alternative results of a challenge depending on certain situations.

2.4 Hybrid Approach

This approach is the combination of the descriptors of statistical, linguistic and machine learning related approach. The graph-related approach is an unsupervised method, where the required words, sentences or paragraphs were ranked based on a graph. Graphical related technique and algebraic functions are employed to isolate the appropriate weight for a statement, thereby obtaining the most relevant word or sentence from a single document. The highest number, K, of the properly weighted statements are chosen for the text short description. In Sankar et al. (2011), a graph-based method was employed to model the text file. Every statement in the text develop into the graph's vertex and the value computed to the vertex. LexRank and TextRank are the best algorithms employed in programme summarisation software using graph method.

Fuzzy Logic: is an approach that mimics the human way of reasoning and making decisions.
The fuzzy logic approach makes use of fuzzy rules and triangular membership functions. The fuzzy rules are in the form of IF-THEN. The

triangular membership function classifies each score into one of these categories such as LOW, MEDIUM & HIGH. Then, fuzzy rules are applied to determine whether sentence is unimportant, average, or important. A fuzzy logic method could look like:

IF F1is Medium and F2 is Medium and F3 is High and F4 is High and F5 is Medium and F6 is High and F7 is High and F8 is High THEN sentence is important and would be added to the summary.

2.5 Transfer Learning Approach

Transfer learning gives machine or deep learning models a capability that employs previous understanding in solving new problems in a new domain. In NLP, transfer learning can be classified in two ways. Transductive transfer learning is a type of learning whereby the source and target task have the same objective. While Cross-Lingual Learning transfers the information to a different language (Sebastian et al., 2017), and Domain Adoption applies the knowledge of the model to another domain. The use of Cross-Lingual Learning solves the problem ofscarcity of dataset on the focused dialect and form machine learning modules for dialect in previous impossible situation.

For inductive transfer learning, the objective of the initial job vary from the final job. In Sequential Transfer Learning, the goal is to train a model on one task and then adopt the model to another task. While in Multi-Task Learning, two or more tasks are jointly learned at the same time. (Ronan & Jason, 2008)

Transfer learning could be described in respect of domain and tasks. A domain (D) comprises of a descriptor space:{\display style {\mathcal {X}}}(X) and a minimal probability distribution P(X) {\display style P(X)}P(, where X = $\{x_1,\ldots,x_n\}$ ϵ X{\display style X=\{x_{1},...,x_{n}\}\in {\mathcal {X}}}X =. Given a specific domain, D = {X, P(X)},

A task is made up of 2 constituents: a label space (Y){\display style {\mathcal {Y}}} and an objective predictive function

$f: X \rightarrow$Y{\display style f:{\mathcal {X}}\right arrow {\mathcal {Y}}}. The function f was employed for the prediction of equivalent label f(x) of a new occurrence x. This task, is denoted by T = {Y, f(x)}, is learned from the training data consisting of pairs $\{x_i, y_i\}$, where $x_i \epsilon$ X and $y_i \epsilon$Y. (Lin & Jung, 2017).

Given a source domain D_S and learning task T_S, and a target domain D_T and learning task T_T, Transfer Learning was targeted towards enhancing the ability to learn of the predictive function of interest, f (t) in Dr by the use of the experience in D_S as well as T_S, where $D_S \neq D_T$, or $T_S \neq T_T$.(Lin & Jung, 2017)

3.0 Automatic Summarization in Seven Languages

3.1 Automatic Summarization of English Texts

In this domain, there are many researches that had been undertaken using several approaches and techniques. However, some significant ones will be highlighted. Programmed summarization software employ groupings based on similarities to produce short description of texts of various titles of files. The text files are interpreted by means of Term Frequency-Inverse Document Frequency (TF-IDF).

The Term Frequency (TF) means average number of appearance of a file in the similarity groupings. The title is represented by words whose TF-IDF importance is greater in the similarity groupings. The choice of applicable statements is centered on the similarity of statements with the title of the similarity groupings. In Zhang and Cun-He (2009), the calculation of resemblance among the sentences is computed based on the resemblance of the words in-between one statement and the other as well as the meaning similarity of words. K-means is then employed to put together the statement of the files in the similarity groupings.

Khushboo (2010) projected a procedure centered on graph algorithm for programmed text summarisation. The procedure is comprised of developing a graph from text. The vertices of the graph stands as the texts statements, for every sentence is represented by a vertice. The side of the graph stands for a connection between the statement, the connection is measured by computing the likeness between the statements.

The weight of each vertice is computed by the use of Cosine function (COS). The short description of text is gotten through choosing of the smallest length from the initial statement of the initial passage, while it ends with final statement.

Widyassari *et al.* (2019)proposed a general structure of Automatic Text Summarization of any language as shown in figure3.1, the input text document of any language is read into the system and the pre-processing commences. The output of pre-processing is loaded into the Processing stage, and then, finally to the post-processing stage where the precise and concise extracted summary is produced.

However, Kumar et al. (2021) proposed a more generic and detailed framework of Automatic Text Summarization of any language as shown in figure 3.2.The proposed framework has three major stages where statements were prepared before use, sorted out and grouped for producing the short text description. Their framework focus on the pre-processing techniques for different languages which includes the tokenization, normalization, stop word elimination and stemming.

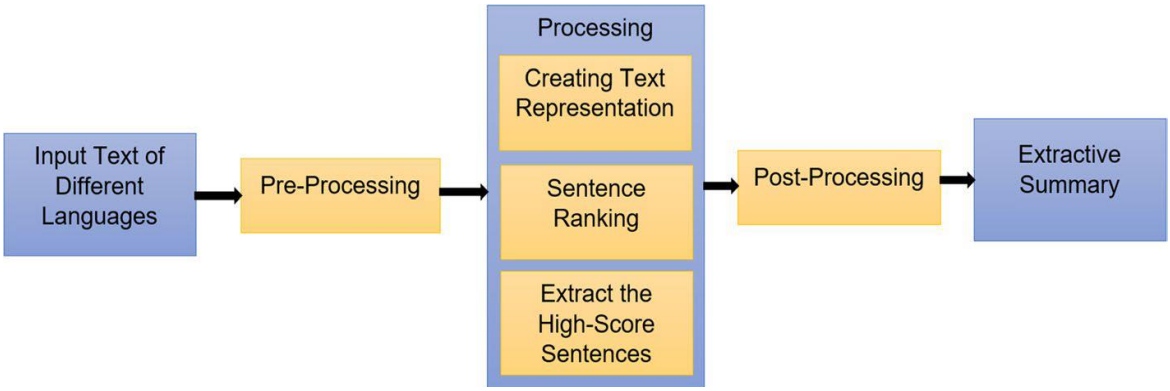

Figure3.1: General structure of Automatic Text Summarization (Widyassari *et al*., 2019)

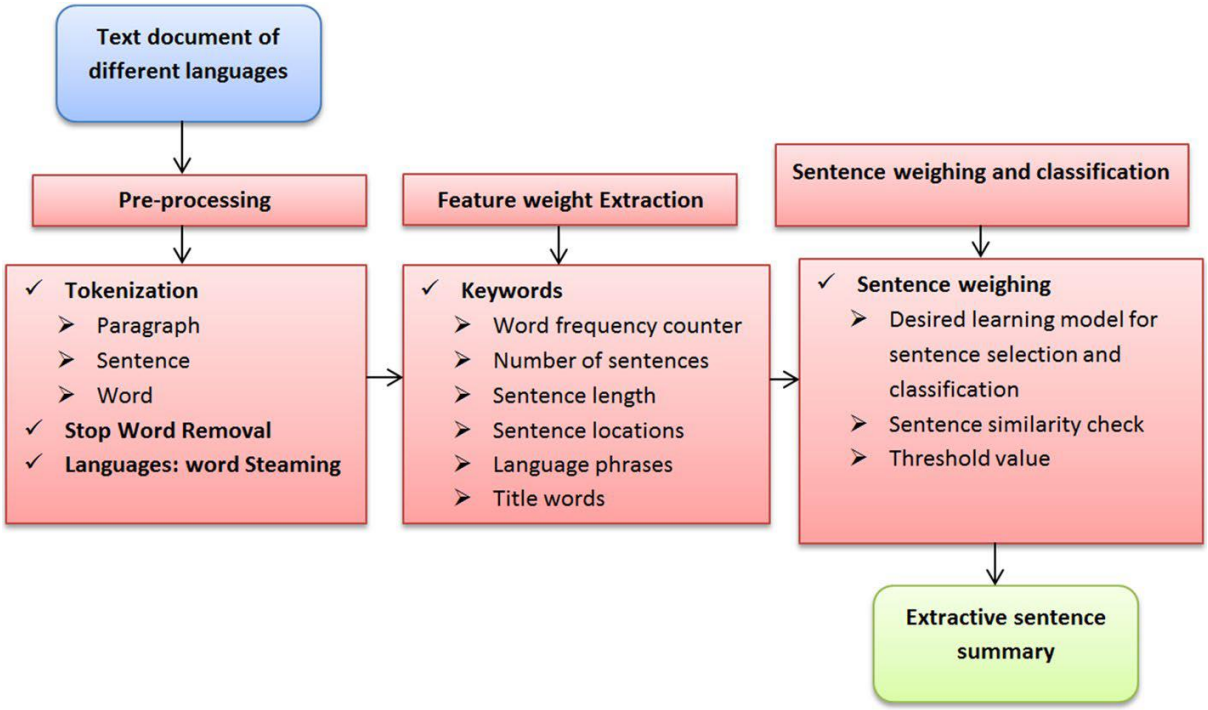

Figure 3.2: General Framework of Automatic Text Summarization (Kumar *et al*., 2019)

Then, the feature weight extraction processes commence where relevant sentences are determined using techniques stated in Figure 3.2 depending on the language structure after which scores were assigned to the attributes by feature-weight formula. Then, eventual weight of individual statements was calculated, best rated statements were employed to put together the final short text description.

Hence, a suitable machine, deep or transfer learning model for sentence selection and classification is applied based on the different language structure. These models are being used to generate extractive summaries by pre-training the models with the suitable predictors that were carefully selected for training the model. The trained model was employed in the process of classification in order to decide on inclusion of a statement in a short text description or not. Finally, the sentence similarity eliminates the identical statements by means of statement similarity subroutine. Just a number of resemblance technique were strongly recognised for the estimation of the extent of resemblance between a number of differing documents. The resemblance between two text files could be estimated by equation 3.1:

$$Distance\left(doc_a doc_b\right) = \left[\left(doc_a - doc_b\right)\left(\left(doc_a - doc_b\right)\right)\right]^{1/} \quad \ldots \quad (3.1)$$

Therefore, the threshold value module is used to determine the statements that finally contributed to the summary and to minimise the generated summary length.

3.2 Automatic Summarization of Chinese Texts

According to Xiaojun and Yuxin (2005), text statements are grouped based on their weights computed based on words count and statement location and then some related statements are set aside as candidate statements, Weight measured by the number of words appearance is computed by the addition of all the words weight in the statements according to TF-IDF procedure. The statements that has been designated as selectable are received by EMD-MMR (Earth Mover's Distance) technique, MMR (Maximum Marginal Relevance) for removing idleness of words. Also, Jiang (2009) proposed that the statistical technique and algorithms might not resolve the difficulty in comprehension of the text file content. The quality of a programmed summarisation centred on main words are enhanced, provided the main words are recognised a-priori. Therefore, the researcher projected a process of determining the main words centred on verbal sequence for producing the programmed summarisation to minimise words repetition. This researcher employs HowNet database in extracting the association among the Chinese words in order to develop the verbal sequence. In enhancing the precision while determining the main words, selected names, action words, word qualifier detected by HowNet plus discovered unfamiliar words which are probably candidate words. After the lexical chains had been established and respective weight determined, a procedure is employed to determine the keywords from the lexical chains.

Deng *et al.* (2020) experimented a double-step Chinese passage short description process employing data on the main words and adversarial learning algorithm. They employed jieba word decomposition software to decompose the headlines and short description information then, apply identifiers determination technique to determine the identifiers from headlines information. They employed very many demonstrations to isolate the best identifier counts as parameters for every statement. The frequency of the identifiers in each statement is initialised to 5, while the magnitude of the terminology, initialised to 60 K. They initialised the proportions of the expressions number in addition to hidden layer to 512. At the translation stage, Adam Optimiser procedure was employed for the fact that greedy procedure straight away select the greatest likelihood result every moment, it does not assure getting the global optimal solution. Hence, beam search procedure, which was initialised to ten, was used to decipher.

The performance evaluation of the model was done using Rouge-1, Rouge-2 and Rouge-L as the assessment technique since Rouge is the most populous evaluation technique in content evaluation domain. Although, rouge assess the short description of text employing the simultaneous appearance of data of a particular number of expressions, it is also a quantification of recall proportion of particular number of expressions.

3,3 Automatic Summarization of Persian Texts

Programmed summarisation of Persians language writings provides novel horizon of research emerging tremendously in recent times. Programmed Persian Text Summariser Zohre and Mehrnoush (2007), employs a fusion technique to programmatically summarise the Persian texts. In this study, lexical chains, graphs centred methods, choice of significant statements built on keywords, like statements frequency, likeness between statements and likeness with the title and user query. PARSUMIST is a different program projected by Shamsfard *et al*. (2009), built on the lexical chains based on enhancement in depiction level, abstract and comprehension of the text by use of synonyms and redundancy detection. PARSUMIST designs have three main divisions: pre-processing, analysis and selection. The systems employ common words (function words), main words and Persian words with same meaning. PARSUMIST do avoid repetition of like statements in the textual short description. Zamanifar and Kashefi (2011) produced Azomto, which programmatically summarise Persian text. It is an ensemble of statistical, mental concept and arrangement based on unique text constituents to be summarised. The projected procedure was employed for the Persian language but applicable to other language too. Following the raw-data transformation stage, Azom follows with development of equivalent text files' self-similar tree in order to determine the main part of documents made up of chapters, part-divisions, passage and expression. Each of the elements of speech to form a statement is checked-up in Persian verbal base in order to isolate the likeness among words.

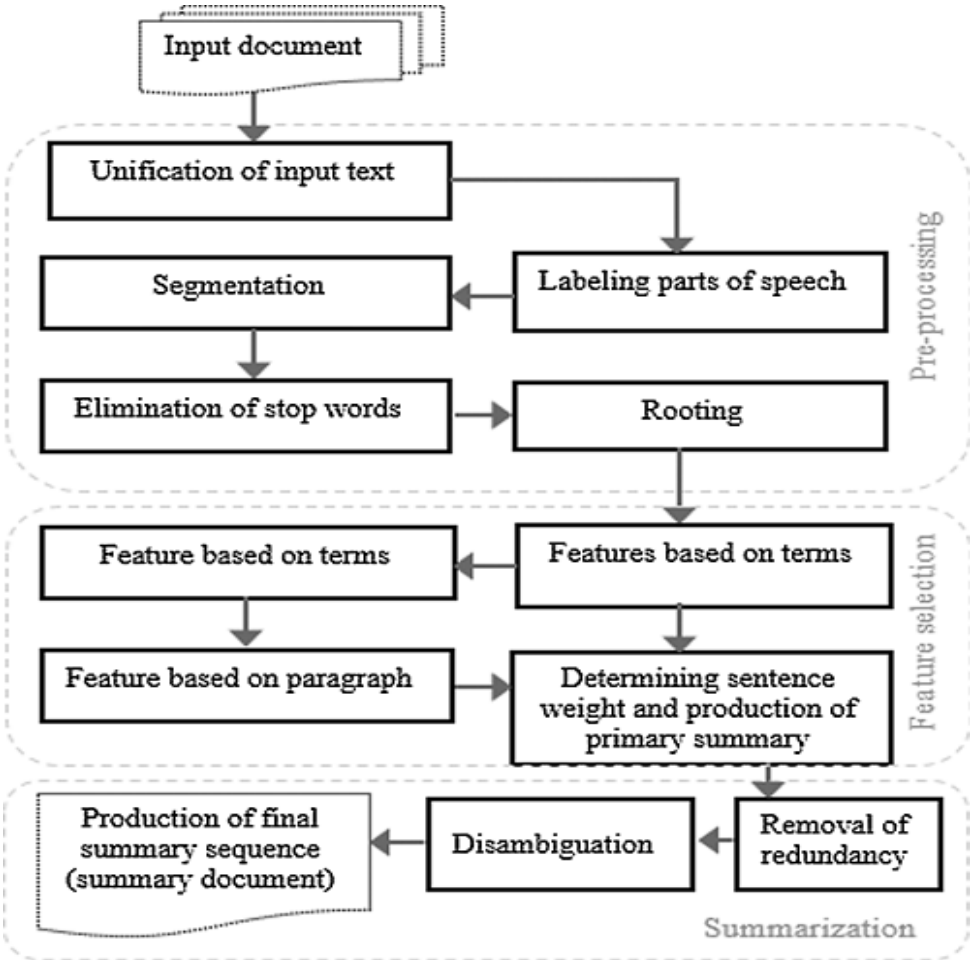

Figure3.3:Automatic Persian Text Summarization Architecture (Heidary et al, 2021).

This paper presents a new combined extractive single-document text summarization method based on text structure. The proposed summarization process consists of three stages: preprocessing, feature selection, and summary generation. This paper presents a new combined extractive single-document text summarization method based on text structure. The proposed summarization process consists of three stages: preprocessing, feature selection, and summary generation.

Heidary *et al*, (2021) proposed a novel extraction based on ensemble single-document passage short description technique that depends on passage organisation. The architecture of the projected procedure comprises of three steps: pre-processing, predictors determination and short description production. In preprocessing stage, they translate the initial passage into a sole whole using a tool called PARSIPARDAZ tool that was produced by the Iranian National Cyberspace Research Institute (Telecommunication Research Center of Iran) in respect of Persian linguistics. The software is 98% accurate in part-of-speech (POS) labelling in addition to being accurate 100% in standardisation (Mazdak & Hassel, 2004).

In the predictors determination step, a predictor analysis is performed on the passage that are fed into the step because the most significant expression that gets to this stage are mostly nouns and adjectives. The most important predictors in the input document are scored and measured in respect of the predictors isolated in the following 3 phases as shown in Figure 3.4.

The essence of the predictors is to assure a correct determination of terms, statement and passages having significant rating in the summary. In the text short description step, the predictors determination steps are germane in the determination of the sentence being selected to be in the short description of text. Each term, sentence or paragraph in the input document has a rank that is determined based on the weight of its features. In addition to term-based and sentence-based features, the feature of paragraphs that consist of sequences of higher importance rating will be selected for the summary.

Then, the weightiness of each statement is determined in respect of the blending of the feature selection factor sequel to the computation of the weightiness of each statement in the data file, the statement are arranged in the order of the highest to the lowest weight. Then, preliminary sets of statements give the text short description, which is later fine-tuned by two procedures, which are redundancy elimination and disambiguation. The procedures produced an optimum text short description through removal of identical and ambiguous statement in the text short description. The evaluation of the efficiency of this projected technique gave 78.5 % precision, 80% recall and 89 b% readers' satisfaction. (Heidary *et al*, 2021).

The concept employed in the projected technique as well as the enhancement applied in the predictors determination step in respect of text arrangement largely minimise the challenges like inconsistency, ambiguity and redundancy in the text short description.

3.4 Automatic Summarization of Arabic Texts

There are few software and researches on programmed text summarization in Arabic language unlike English and French. El-Shishtawy and El-Ghannam (2012) projected concerning Arabic programmed text short description based on exaction, derivation and sorting of Arabic elements of a statement were employed for this study for computing the unique attributes. The evaluation of the significance of a statement was done by selecting the main phrases. Aside of employing the terms frequency and terms distance, the extractor used the linguistic knowledge for enhancing its effectiveness. Douzidia and

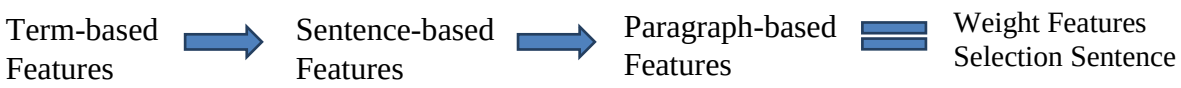

Figure3.4: Feature Selection Stage in Persian Text Summarization (Heidary *et al*, 2021).

Lapalme (2004) produced Lakhas system centred on normalisation by substituting various characters alternatives with a single one, stop words elimination and lemmatisation. The number of words in a statement, suggestive communication, statement location with the TF-IDF importance of every elements of a statement were employed to compute each sentences weight.

Haboush and Al-Zoubi (2012) described a system that determined the parent of each element in a statement. The words can be classified into different similarity groups centred on text roots. The assumptions about the system is that, significant words in the text are present many times. Hence, the germane characteristics of concern for Arabic text summarisation are words frequency and suggestive expressions to enhance a sentence's prominence.

Elbarougy *et al*. (2020) presented a method in Arabic text short description. Arabic linguistic is complicated morphologically making it challenging to filter nouns that should be employed for text short description procedure. So, Al-Khalil morphological analyser is employed to resolve the difficulty in noun filtering. The projected technique is graph-based, which possess the text file as a graph and the vertices represent the statement. An Enhanced PageRank procedure is employed with the first score for every node, which is frequency of nouns in a particular statement. Text summarisation procedure comprises of three main steps: pre-processing step, features selection and graph building step as well as employing the tweaked PageRank algorithm and summary drafting, Figure 3.5.

To assess the efficiency of the technique, Essex Arabic SummariesCorpus (EASC)was employed by default and its precision, recall and F-measure were computed. LexRank and TextRank procedures were employed under same conditions, the projected technique gives better performance in comparison with other Arabic text short description methods. The performance is better because Alkhalil morphological analyser, whichw as employed to minimise the difficulty in Arabic structure complication and gives finest investigation.

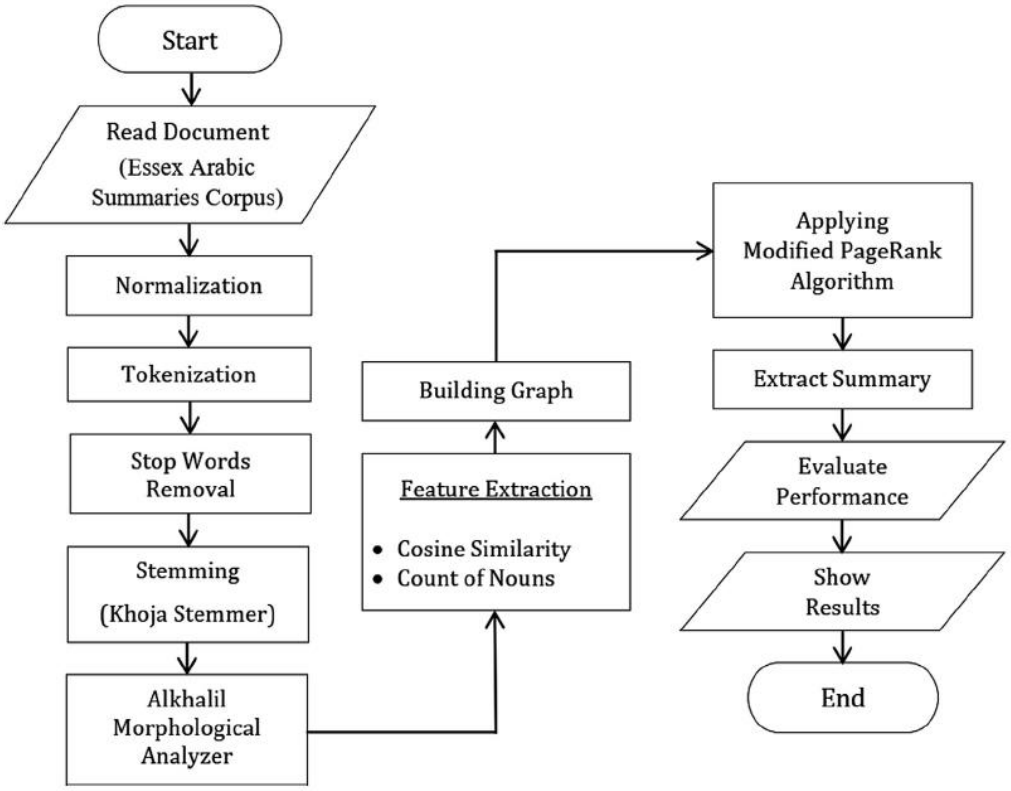

Figure3.5:Text Summarization of Arabic Language (Elbarougy, et al., 2020)

3.5 Automatic Summarization of of Spanish Texts

In this summarization process, Esteban and Lloret (2017) employed the use of some natural language processing techniques and tools for retrieving and extracting information, as well as carrying out a linguistic analysis of the documents. The whole process begins from the document extraction phase which was gotten from twitter and travel website. The Document filtering phase goal is to discard information that is not in Spanish or that does not give an opinion about the research domain. The Sentence segmentation splits the Spanish filtered documents into sentences and classify the sentences into their sentiment (neutral, positive, and negative). The Sentence grouping categorizes similar sentences to avoid redundancy in the summaries, and this is done using Cosine metric. The whole process to create the summaries is depicted in Figure 3.6:

In the Summary generation phase, the sentences were ranked by groups and subsequently, they improve the summaries coherence and readability by employing some techniques and rules to impersonalize the summary to an extent. Finally, this generated three types of summaries, a mixed summary that shows good and bad, a positive summary with the best aspects, and a negative summary that highlights the worst aspects. The summary model was evaluated using a user evaluation method using a questionnaire, wherein, 41 people answered and gave their opinions about the summary generated. The results of the evaluation showed that the summarization service is very useful, with 75% accuracy.

3.6 Automatic Summarization of Hausa Texts

Muazzam *et al*. (2017) conducted a research to develop a model to automatically summarize Hausa Language text based on feature extraction using Naive Bayes model. They adopted five features such as keywords, title, cue phrases, sentence length and location in the summarization process. Then, Naive Bayes model is used to weigh each sentence based on its features. The automatic text summarization tested on the Hausa Language dataset is better if morphologic analysis is considered. The figure 3.7 shows the design of this summarization system. and its 3 stages which include pre-processing, feature weight extraction, and sentence weighing and simplification.

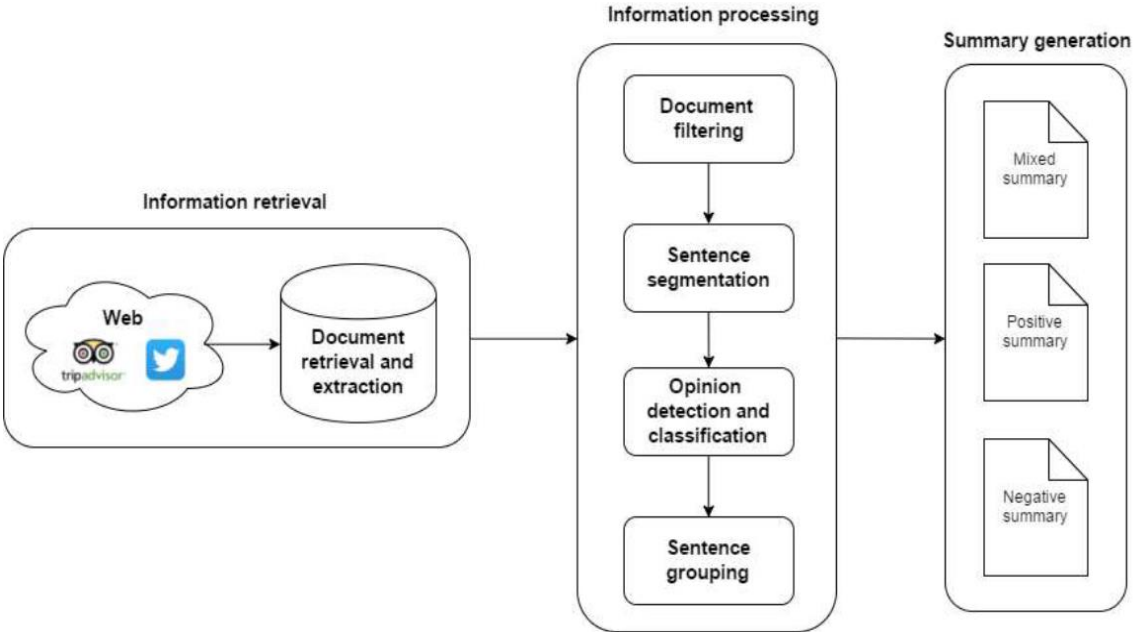

Figure 3.6: Overview of the Spanish text summarization approach (Esteban and Lloret,2017).

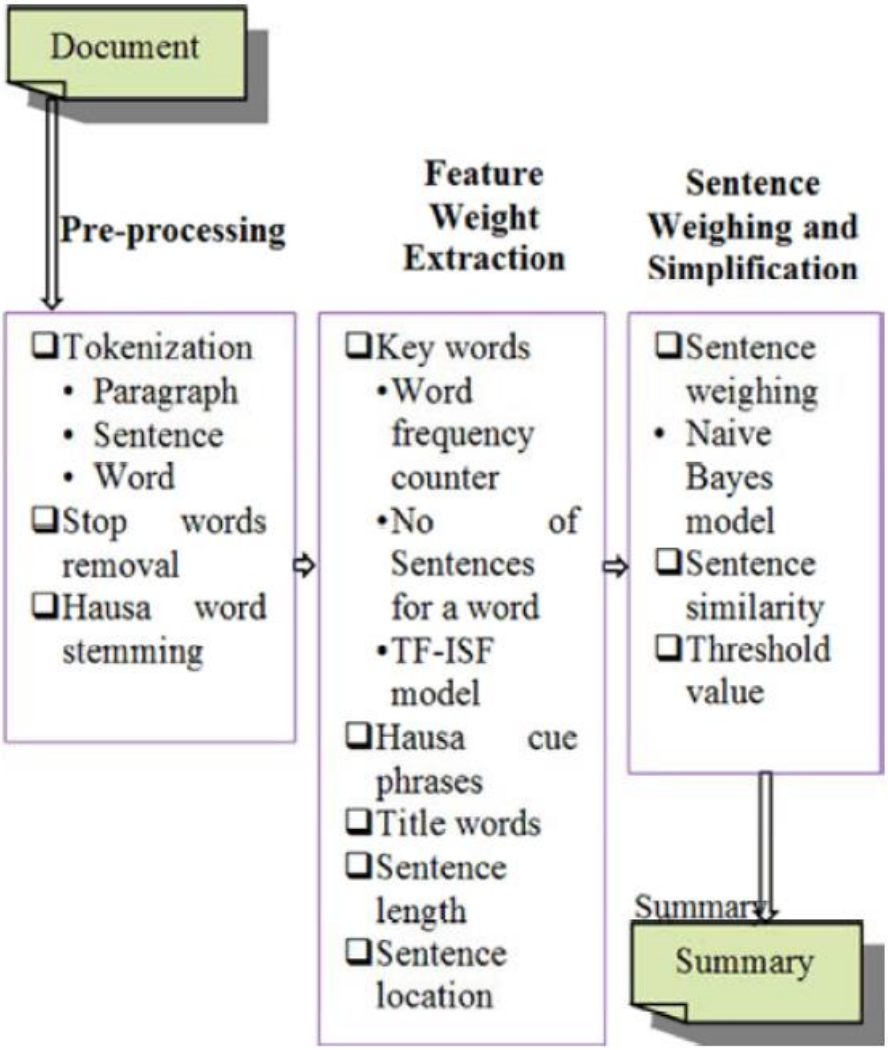

Figure 3.7: Hausa Language text summarizer model (Muazzam et. al, 2017)

The system reads an input text file and uses tokenization to break down the document into groups of smaller units, namely: paragraph, sentence, and word units. The stop words removal eliminates the frequently occurring words that do not have any meaning like prepositions and articles such as "on" and "the". The resulting content words are stemmed using the Hausa word stemming algorithm developed by (Bashir et al., 2015).

In the feature extraction stage, the weight of each word is determined using TF-ISF model. Meanwhile, a set of keywords is formed by choosing the top 10 words with the highest TF-ISF values. The system assesses each sentence using five modules such as Hausa cue phrase, keywords, title words, sentence length and location modules.

In the final stage, the system uses Naive Bayes model to determine a total weight for each sentence using the five different values. A sentence similarity module is used to eliminate any duplicate sentence. A threshold value module is used to choose the summary sentences. The model is evaluated using the precision, recall and F-score values of the summary based on the 3 corresponding experts' summaries.

4.0 PROPOSED AUTOMATIC SUMMARIZATION OF YORUBA TEXTS

4.1 A Brief Review on Yoruba Language

Yoruba is a tonal language of the Niger-Congo family whose many varieties are spoken across West Africa by about thirty million people. The linguistic is used for communication in other

communities such as Brazil, Cuba, Haiti, Togo, Benin Republic, Trinidad and Tobago. Yoruba is a language, out of the three main linguistics used as a medium of communication in Nigeria especially, in the South Western Nigeria. The Yoruba dialect consists of several dialects such as Egba, Ibadan, Ijesha, Ife, Ikale, Oyo and Shaki, which is spoken across several region in the South-Western part of Nigeria. The Ibadan and Oyo dialects are considered as the standard Yoruba for Yoruba literacy (Carter-Enyi, 2018).

Yoruba has a 3 flat pitches and 2 or 3 line pitches. Each unit of pronunciation has a minimum of a pitch. Although, a unit of pronunciation having one lengthy open vocal tract speech sound may comprise of two pitches. Pitches differentiate by means of sharp intonation for high pitch (⟨á⟩, critical intonation belonging to low pitch (⟨à⟩, the normal pitch is perfect, except relating to a syllable that indicate it by a bar over a letter (⟨a⟩, ⟨n̄⟩) (Carter-Enyi, 2018). When instructing on reading and writing in Yoruba, sol-fa syllables of soft notes are employed to identify the pitches: low is "do", normal is "re" and high is "mi" (Carter-Enyi, 2018). It follows the basic order of subject–verb–object pattern which is the same with English Language and some other languages of the world.

4.2 Discussion of the Proposed Text Summarization Tool in Yoruba Language

In this approach, the implementation of the text short description software shall be factored into three stages, which comprises of: pre-processing, feature-extraction and summary generation step. In figure 3.8, the system will read a Yoruba document. This document will be broken into several paragraphs using a Paragraph Segmentation module. In the Sentence Normalization module, punctuations, special characters, and digits will be eliminated from the sentence. The Tokenization module ensures that the paragraphs are broken into statements and eventually, the statements into expressions. Hence, the Stop Word Filtering module, removes the delimiters and confines the passage into its summarised form. Since Yoruba is an SVO (Subject-Verb-Object) language (Aladesote *et al*., 2011), the Yoruba Morphology Lexical Analyser module developed by Aladesote *et. al*, (2011)will process every sentence to a Subject-Verb-Object pattern. Then, each expression in the statement picks a tag denoting its part of speech (POS) location in a statement, which is done by the Part of Speech Tagging module developed by Adedjouma *et al.(*2013). The location of the expression could denote it in noun, action word, preposition and delimiter article. The procedure could be employed to find the frequency of nouns in every statement.

The feature extraction processes commence by using the Keyword Frequency which checks for the relevance of each word in the document by counting how many times it occurred in the document. The keyword with the highest frequency is likely to be present in the generated summary. The Sentence Position implies that sentences in a specific position determines their relevance to the subject context, the important sentences will be presented in the first or last position. Finally, the Sentence Length determines the relevance of a sentence based on its length. Most short and long sentences have small values as they are not suitable for the summary. A Cross-Lingual model, which is a Transfer Learning model, will be used to develop the Yoruba text summarizer. Then, an evaluation will be carried out to compare the extracted and summarized samples technique(s) with those summarized by the self-developed model using Precision, Recall and F-measure as metrics.

This means terms that have occurred over and over and that increase the score of their sentences. It reflects how important the word is for the document.

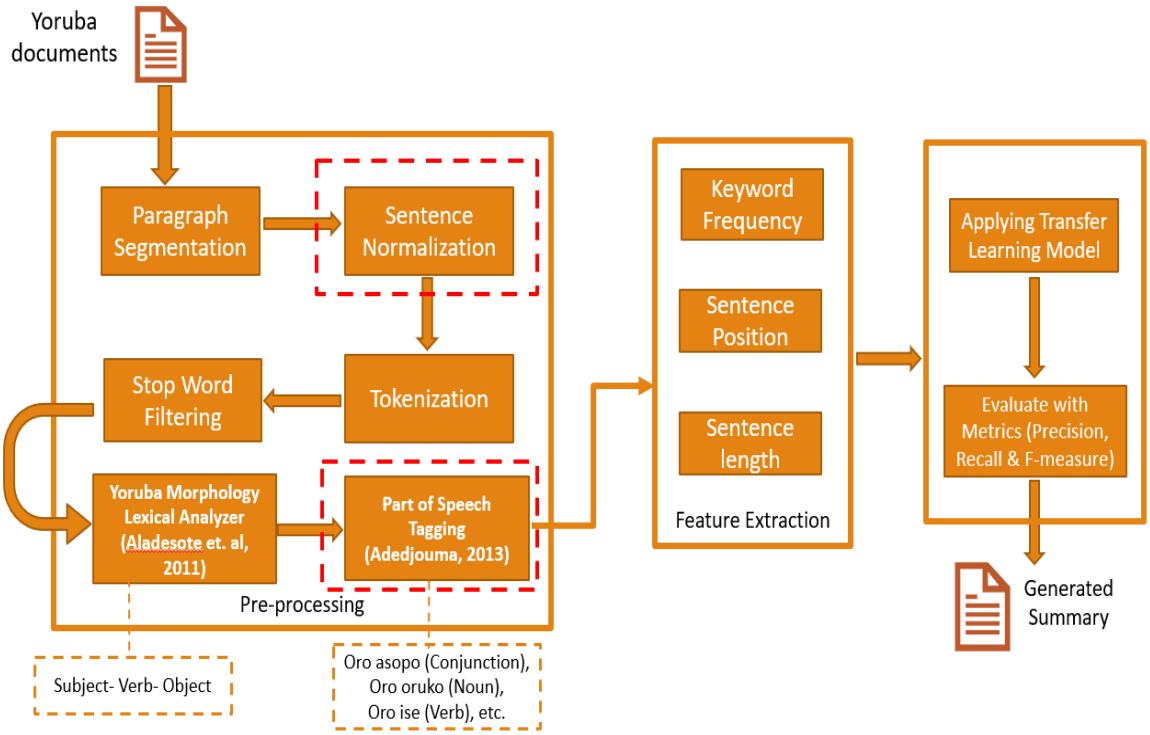

Figure 3.8: Proposed Methodology of Yoruba Text Summarization approach

4.3 Challenges Encountered in Automatic Text Summarization of Other Languages

This section discussed some of the problems encountered while researching about different languages and their impacts on the quality of summaries generated. The Chinese language belongs to the CJK language family (Chinese, Japanese, and Korean). These languages do not use spaces or punctuations to separate words in sentences which makes it extremely difficult to tokenize or segments words in the sentence.Before applying any automatic summarization technique, the lexis and semantics of the language must be critically analysed with linguistics experts of the language. The lexical and semantic analysis of each language are different across languages based on the difficulty and uniqueness of every language and the level of research by other authors on that language. The worst challenge in the discipline of information Technology is summarisation of text using effective algorithm and procedure.

In Chinese language, the main problem encountered is the text segmentation. This has to do with isolating the characters that makes up words and the delineation among the words in their normal language settings. It is as well employed to separate a text into paragraphs, statements and words thereby, identifying the word syllables. In English, like Yoruba and other languages, the separation of texts is easy because spaces and punctuations can be employed as dividers in text separation into words. Nevertheless, several Asian languages like Chinese, Japanese, Korean, Khmer (Cambodian) do not separate the words by spaces or punctuations and does not have a fixed length of characters. Also, it is very difficult to identify keywords in Chinese texts because most Chinese characters may be a word or a part of a word without space or punctuation identifying word demarcations. A way to resolve these problems, different Chinese text segmentation algorithms (N-grams algorithm, Dictionary-based algorithm, Manticore Search, Character-net based algorithm, Conditional random fields) were projected years ago but none is yet to be accepted as a standard. Also, there have been a very active research around keywords identification for Chinese language but there is not a commonly accepted one because of the several kinds of keywords identification methodology in use today.

In the Persian text, one of the major issues is the Ezafe marker, which is an unstressed vowel placed in-between prepositions, nouns or adjectives in expressions verbally pronounced but not written so, which causes problems of syntax and meanings. In the pre-processing stage, both the Persian and Arabic languages have a complex morphology because of the variations in their word forms, Persian stemming procedure is used to minimise infested words to the form of the root. Several words in Persian and Arabic languages are root-derived consisting of 3-charcaters.

Furthermore, the Persian text lacks linguistic resources such as semantic dictionaries, stop and keyword lists, computational dictionaries, thesaurus, corpus, term ontology, semantic vocabulary and lexical ontology such as WordNet in English are important resources for generating an efficient Persian text summarizer. Unavailability of resources of the like and language handling facilities facilitates structural generalisation to the Persian language a difficult tasks. Although, efforts have been made to generate the necessary resources such as the Persian richest lexicon which is made by Eslami (2004), known as, FarsiNet or Farsi, WordNet remains solely the resource for lexical semantics.

On Arabic text, there exist only a number of researches in Arabic NLP because Arabic language is complex due to complex morphology: word form variations and several written forms that letters arrangements can attain. Ambiguity in Arabic words is not unconnected with its three-consonantal root. In order to resolve the problem of disambiguation of words, words formation and identification of the internal structure of words, a linguistic morphology analysis was introduced to ARAMORPH dealing with only Arabic dictionaries words and MORPH-2, a lexicon-based morphological analysis software containing 3266 root words including related properties (Alami *et al.*, 2015). Moreover, one of the most challenging issues in Arabic language is the word stemming because many Arabic words are derived by adding prefixes and suffixes to the original root words. Therefore, extracting lemma, stem or root is a hard problem for Arabic language.

Isolating the root of Arabic word stemming facilitates the one-to-one mapping of grammatical variations of word to instances of same term. Alami *et al*. (2015) improved the integrity of Arabic text summarisation employing statistical feature selection, structural and conceptual methods.

5.0 CONCLUSION

This study proposed a methodology for extractive text summarization of Yoruba Language. It aimed at bridging the research gaps in developing a text summarization tool for Yoruba Language. Although, there are few existing researches on developing resources like standard datasets, stop word lists, part of speech taggers for Yoruba language like in English Language but Yoruba Language as one of the three major languages spoken in Nigeria, majorly by the South-Western part of Nigeria and some communities in other countries like Brazil, Cuba, Haiti, Togo, Benin Republic, Trinidad and Tobago, has no research about Yoruba text summarization, as far as we know. Furthermore, Yoruba language has been classified as a language in serious danger of extinction by UNESCO Red Book on endangered languages (Abiola, 2020). Therefore, Yoruba language deserves much more attention from scientists because there are not many works in the field of natural language processing for Yoruba.Due to this research gap, this study proposed to develop and implement a technique for automatic text summarization in Yoruba language.

6.0 REFERENCES

Abiola, O. B. (2020). A Web-Based Yorùbá to English Bilingual Lexicon for Building Technicians. *International Journal of Advanced Trends in Computer Science and Engineering*, vol, 9, pp, 793-800. DOI:10.30534/ijatcse/2020/114912020.

Adejauma, S., Aoga, J., Igue, M, A., (2013) Part-of-speech tagging of Yoruba standard, language of Niger-Congo family. In: *Research Journal of Computer and Information Technology Sciences*, Vol. 1, no.1, pp. 2-5. http://hdl.handle.net/2078/211882

Aladesote, I, Olaseni O, E, Adetunmbi, A, O,& Akinbohun, F, (2011). A Computational Model

of Yoruba Morphology Lexical Analyzer. *International Journal of Computational Linguistics* (IJCL), Vol. 2, Issue 2, pp 37 – 47

Alami, N, Meknassi, M,& Rais, N, (2015). Automatic Texts Summarization: Current State of the Art. *Journal of Asian Scientific Research.* Vol, 5, Pp, 1-15. DOI:10.18488/journal.2/2015.5.1/2.1.1.15.

Bashir, M., Rozaimee, A., B., & Isa, W., M., (2015). A word stemming algorithm for Hausa language, *IOSR Journal of Computer Engineering*, 17: 25-31, pp 68-73.

Carter-Énylì, A. (2018). Hooked on Sol-Fa: The do-re-mi heuristic for Yorùbá speech tones. *Africa*, 88(2), pp, 267-290. doi:10.1017/S0001972017000912

Deng, Z, Ma, F, Lan, R, Huang, W,& Luo, X, (2020). A Two-stage Chinese Text Summarization Algorithm Using Keyword Information and Adversarial Learning. *Neurocomputing*. 425, DOI:10.1016/j.neucom.2020.02.102.

Elbarougy, R, Behery, G, &El Khatib, A.,(2020). Extractive Arabic Text Summarization Using Modified Page Rank Algorithm, *Egyptian Informatics Journal* vol, 21, pp, 73–81, https://doi.org/10.1016/j.eij.2019.11.001

El-Shishtawy, T., & El-Ghannam, F., (2012). "Key phrase based Arabic summarizer (KPAS),",presented at Informatics and Systems (INFOS), *8th International Conference*. NLP-7, NLP-14.

Eslami, M., (2004). "Persian generative lexicon," in Proceedings of The 1st Workshop on Persian Language and Computer, University of Tehran, with the Cooperation of the Research Center of Intelligent Signal Processing (RCISP), Tehran, Iran.

Esteban, A., & Lloret, E., (2017). TravelSum: A Spanish Summarization Application focused on the Tourism Sector Proces amiento del Lenguaje Natural, Vol., 59, pp. 159-162

Douzidia, F, S,& Lapalme, G, (2004). "Lakhas, an Arabic summarization system," in *Proceedings of 2004 Document Understanding Conference (DUC2004)*, Boston, MA.

Haboush, A.,& Al-Zoubi, M., (2012). "Arabic text summarization model using clustering techniques," *In World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741*, vol. 2, pp. 62–67.

Heidary, E, Parvin, H, Nejatian, S, Bagherifard, K,& Rezaie, V, (2021). Automatic Persian Text Summarization Using Linguistic Features from Text Structure Analysis. Computers, Materials & Continua. Vol, 69, pp, 2845-2861, DOI:10.32604/cmc.2021.014361.

Jiang, X, Y, (2009). "Chinese automatic text summarization based on keyword extraction," in *Database Technology and Applications, First International Workshop*, pp. 225-228.

Kumar, Y, Kaur, K,& Kaur, S, (2021). Study of automatic text summarization approaches in different languages. Artificial Intelligence Review. DOI:54. 10.1007/s10462-021-09964-4.

Khushboo, S, T, Dharaskar, R, V, D,& Chandak, M, B, (2010). "Graph-based algorithms for text summarization," presented at the Third International Conference on Emerging Trends in Engineering and Technology

Lin, Y, P, & Jung, T, P. (2017) Improving EEG-Based Emotion Classification Using Conditional Transfer Learning. Front. Hum. Neurosci. Pp, 11:334. doi:10.3389/fnhum.2017
.00334

Mazdak, N,& Hassel, M.,(2004). "FarsiSum-A Persian text summarization," Sweden: Stockholm University,Department of linguistics, Master Thesis,

Muazzam, B., Azilawati, R., & Wan-Malini, W., (2017). Automatic Hausa LanguageText Summarization Basedon Feature Extraction using Naïve Bayes Model, *World Applied Sciences Journal* 35 (9): 2074-2080,DOI: 10.5829/idosi.wasj

Pedersen, T, (2010). Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text. Human Language Technologies. The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, pps 329–332

Sankar, K., Vijay, S., Ram, R., & Sobha, L. D., (2011) Text Extraction for an Agglutinative Language, Problems of Parsing in Indian Languages, Special Volume, PP 56-59.

Shamsfard, M., Akhavan, T., & Jourabchi, M., E., (2009). "Parsumist: A Persian text summarizer," Natural Language Processing and Knowledge Engineering, NLP-KE, International Conference.

Widyassari, P, A, Affandy, N, E, Fanani A, Z, Syukur, A, & Basuki, R, S, (2019) Literature review of automatic text summarization: research trend, dataset and method. In: *IEEE International conference on information and communications technology (ICOIACT*), pp 491–496.

Xiaojun, W, & Yuxin, P, (2005). "A new re-ranking method for generic Chinese text summarization and its evaluation," in Proceeding of: Digital Libraries: Implementing Strategies and Sharing Experiences, *8th International Conference on Asian Digital Libraries*, ICADL, Bangkok, Thailand. Proceedings. Springer Berlin Heidelberg, pp. 171-175.

Zamanifar, A., &Kashefi, O., (2011). "AZOM: A Persian structured text summarizer," in *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems* (NLDB'11). Springer-Verlag Berlin, Heidelberg, pp. 234-237.

Zhang, P, Y, & Cun-He, L, (2009) "Automatic text summarization based on sentences clustering and extraction," Computer Science and Information Technology, ICCSIT, 2nd IEEE International Conference.

Zohre, K., & Mehrnoush, S., (2007). "A system for automatic Persian text summarization," presented at *the 12th International CSI Computer Conference.*