# AN EFFICIENT FEATURE SELECTION MODEL FOR IGBO TEXT

Ifeanyi-Reuben Nkechi J.[1] and Benson-Emenike Mercy E.[2]

[1]Department of Computer Science, Rhema University Nigeria
[2] Department of Computer Science, Abia State Polytechnic Nigeria

## *ABSTRACT*

*The development in Information Technology (IT) has encouraged the use of Igbo Language in text creation, online news reporting, online searching and articles publications. As the information stored in text format of this language is increasing, there is need for an intelligent text-based system for proper management of the data. The selection of optimal set of features for processing plays vital roles in text-based system. This paper analyzed the structure of Igbo text and designed an efficient feature selection model for an intelligent Igbo text-based system. It adopted Mean TF-IDF measure to select most relevant features on Igbo text documents represented with two word-based n-gram text representation (unigram and bigram) models. The model is designed with Object-Oriented Methodology and implemented with Python programming language with tools from Natural Language Toolkits (NLTK). The result shows that bigram represented text gives more relevant features based on the language semantics.*

## *KEYWORDS*

Feature Selection, Igbo Language, Igbo Text Pre-Processing, Text Representation

## 1. INTRODUCTION

As the world advances in Information Technology together with huge amount of textual data it generates, an intelligent text-based system becomes the remedy for effective management of these data. The selection of optimal set of features from these texts is necessary in order to boast the performance of the system. According to Pradnya and Manisha, 90% of the world's data is unstructured (textual data) and there is necessity for intelligent text analysis on the data [1]. The major challenge of intelligent text-based system is accuracy of the system and high dimensionality of feature space [2]. It is very important to use feature selection model to capture these challenges, by reducing high dimensionality of data for effective text-based system. The selection of irrelevant and inappropriate features may confuse the system and can lead to incorrect results. The major problem of text-based system is the enormous quantity of features which reduces system performance and consumes higher time [3]. Noura et al. defined feature selection, as the process of choosing important features for use in text model construction to improve the performance of the model [4]. Feature selection process is highly recommended in any text-based system to select the most relevant features, thereby reducing the feature dimension space and improving the system performance.

Ladha and Deepa emphasized feature selection as a subset selection process mainly employed in machine learning, whereby subsets of the features available from the datasets are selected for application of a learning algorithm [5]. In data mining, before applying any mining technique, the irrelevant attributes needs to be filtered and the filtering can be done using different feature selection techniques such as wrapper, filter and embedded techniques [6]. Feature selection is a research domain surrounding text mining, data mining, Natural Language Processing (NLP) and Machine Learning.

DOI: 10.5121/ijdkp.2018.8602 

The advancement of Information Technology has motivated the use of Igbo language in the creation of resources, articles publications and news reports online [7]. As the number of Igbo texts online and tasks that necessitate feature selection on Igbo text documents are increasing, there is need to have an effective model to select most relevant features on Igbo text corpus to improve the performance of any system with the text.

This paper analyzed the structure of Igbo text and designed an efficient feature selection model for an intelligent Igbo text-based system. This is to enhance the robustness of the system by removing the noisy and redundant features from the features space.

## 2. IGBO LANGUAGE STRUCTURE

Igbo is one of the three main languages (Hausa, Yoruba and Igbo) in Nigeria. It is largely spoken by the people in the eastern part of Nigeria. Igbo language has many dialects. The standard Igbo is adopted formally for this work. The present Igbo orthography (Onwu Orthography, 1961) is based on the Standard Igbo. Orthography is a method of writing sentence or building grammar in a language. Standard Igbo has thirty-six (36) alphabets (a, b, ch, d, e, f, g, gb, gh, gw, h, i, ị, j, k, kw, kp, l, m, n, ṅ, nw, ny, o, ọ, p, r, s, sh, t, u, ụ, v, w, y, z) [7].

## 3. FEATURE SELECTION METHODS

Feature selection methods are categorized into three types. These are Filter methods, Wrapper methods, and Embedded methods [1] [6].

### 3.1 FILTER METHOD

In the filter method, the feature selection is independent of the learning algorithm to be applied to the selected features and assesses the relevance of features by considering the principle criteria in ranking technique adopted. The features are given a score using a fitting ranking criterion and the features having score below some threshold value are considered irrelevant and are removed [1]. The basic filter feature selection algorithms are Chi-square statistics, Information Gain (IG), Document Frequency, Mean Term Frequency-Inverse Document Frequency, Entropy-based (En) and Chi-Square variant [8].

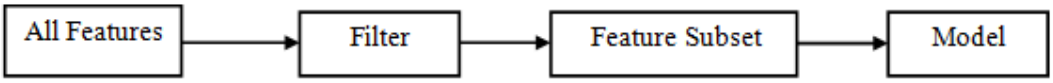


Figure 1. Filter Method

Figure 1 is a sketch of filter method approach of feature selection. Filter method provides a set of most relevant features as the resulting subset not minding the adopted model. Filter approach is a sequential process, not model-oriented and has low cost of computation.

### 3.2 WRAPPER METHOD

In the wrapper method, the feature selection model uses the result of the learning algorithm to determine how good a given feature data subset is. A search process is defined in the space of possible feature subsets. Then various subsets of features are generated and evaluated. This implies that the quality of an attribute subset is directly calculated by the performance of the learning algorithm applied to that attribute subset. Wrapper approach is more computational expensive because the learning algorithm is applied to each feature data subset considered by the search [6]. The wrapper feature selection algorithms are Sequential selection algorithm and Heuristic Search Algorithms.

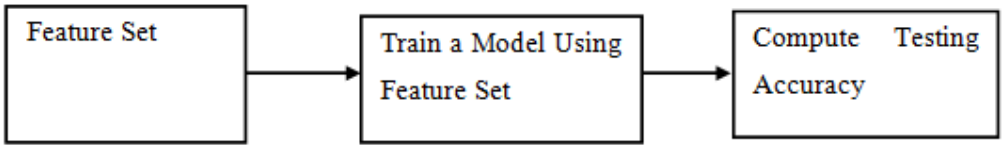


Figure 2. Wrapper Method

Figure 2 describes the selection processes of a wrapper method. The wrapper method is model-oriented and normally gets good performance based on the chosen model. It is an iterative process and has high computational cost. In each iteration, several subsets of input features are generated and the accuracy is tested on the employed model. The subsets of the features that will be tested in the next iteration are based on success of the model for individual feature subsets.

### 3.3 EMBEDDED METHOD

In embedded approach, the feature selection method is built into the learning system algorithm (figure 3). Pradnya and Manisha defined embedded method as a hybrid model because it is a combination of filter and wrapper methods [1]. Some of the examples of embedded method of feature selection are Decision trees, Naive Bayes and Support Vector Machine (SVM).

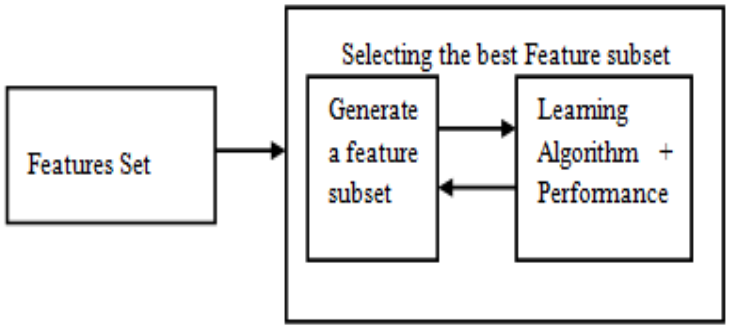


Figure 3. Embedded Method

Mean Term Frequency-Inverse Document Frequency (Mean TF-IDF) feature selection model is used in this work. This is adopted to choose the most relevant features to boost performance and decrease memory cost by reducing the feature dimension space.

## 4. RELATED WORKS

Feature selection is an active research area and various algorithms have been executed to solve feature selection problems. Majority of the works were carried out for English, Spanish, Arabic, French, and Turkish texts but little or no work has been done for Igbo text. Some of the papers related to the work were studied and are presented below:

An effective feature selection method for improving the performance of Arabic text classification was carried out in [9]. The experiments confirmed their proposed model outperforms the existing methods in terms of accuracy.

Aisha et al. examined the performance of five mostly used feature selection methods (Chi-square, Correlation, GSS Coefficient, Information Gain and Relief F) on Arabic text classification and initiated an approach of combination of feature selection methods based on the average weight of the features. The results of experiments carried out using Naïve Bayes and Support Vector

Machine classification model show that finest classification results were obtained when feature selection is done using Information Gain method. The results also prove that the combination of numerous feature selection methods outperforms the finest results gotten by the individual methods [3].

Pradnya and Manisha surveyed three feature selection techniques (filter, wrapper and embedded) and their effects on text classification. The survey proved that filter method should be adopted if the result is needed in less time and for large dataset; and wrapper method should be adopted if the accurate and optimal result is needed. It was also observed that the performance of different algorithms differs according to the data collection and desires [1].

Ifeanyi-Reuben et al. presented a paper on the analysis and representation of Igbo text document for a text-based intelligent system. The result obtained in their experiment showed that bigram and trigram n-gram model gives an ideal representation of Igbo text document for processing text analysis, considering the compounding nature of the language [7].

The performance of various text classification systems in different cases using feature selection with stemming and without stemming on Arabic dataset was compared in [10]. Various text classification algorithms such as Decision Tree (D.T), *K*-Nearest Neighbors (*K*NN), Naïve Bayesian (NB) Method and Naïve Bayes Multinomial (NBM) classifier were adopted. The result showed the classification accuracy for Decision Tree, Naïve Bayesian method and Naïve Bayes multinomial is better than K-Nearest Neighbours (KNN) in all tested cases.

Rehab et al. also presented and compared three feature reduction techniques on Arabic text. The techniques include stemming, light stemming and word clusters. The effects of the listed techniques were studied and analyzed on the Arabic text classification system using K-Nearest Neighbour algorithm. The result from the experiments shows that stemming reduces vector sizes, and hence enhances the classification accuracy in terms of precision and recall [11].

Mohammad et al. proposed a feature selection model-based ensemble rule classifier method for a classification exercise. The experiment was performed on public real dataset. The result shows optimal set of attributes is realized by adopting ensemble rule classification method, as well as the significant improvement in accuracy [12].

Sabu M.K. presented a novel hybrid feature selection model by integrating a popular Rough Set based feature ranking process with a modified backward feature elimination algorithm, to predict the learning disability in a cost effective means. The experimental results show the proposed feature selection approach outperforms the other approaches in terms of the data reduction and system performance [13].

## 5. MATERIALS AND METHODS

The architectural design of the feature selection system for Igbo Intelligent text-based system together with all its tasks is shown in figure 4. This system serves as a filter to mute out irrelevant, unneeded and redundant attributes / features from a given collection of Igbo textual data. This is necessary to boost the performance of any intelligent text-based system in Igbo when incorporated. Improving the feature selection will surely improve the system performance. The tasks in this system are:

1. Igbo Textual Documents Collection
2. Igbo Text Pre-processing
3. Feature Representation
4. Feature Selection

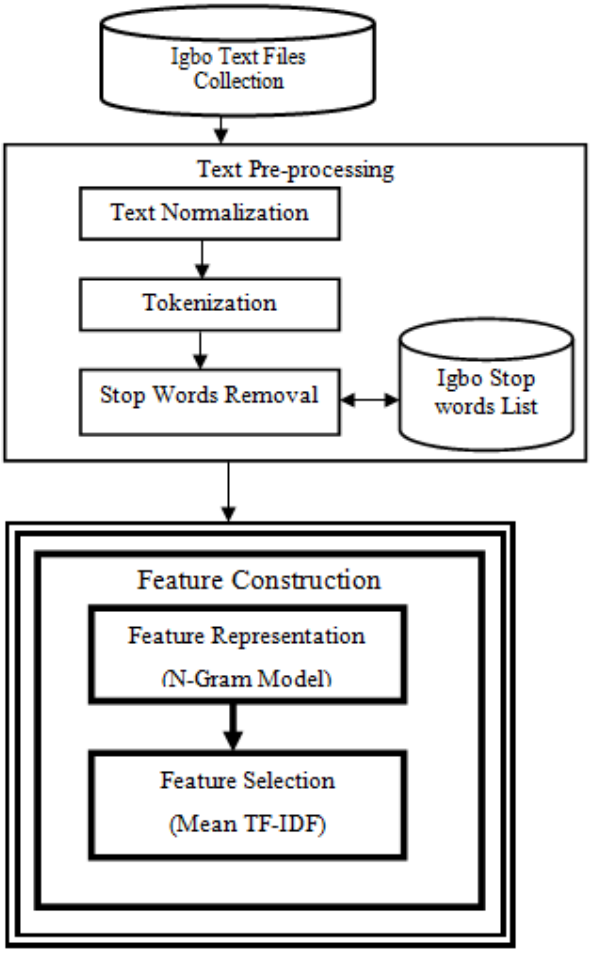


Figure 4. Architecture of Igbo Feature Selection Model

### 5.1 IGBO TEXT COLLECTIONS

The operation of any text-based system starts with the collection of textual data documents. The Unicode model was used for extracting and processing Igbo texts from file because it is one of the languages that employ non-ASCII character sets like English. Processing Igbo text needs UTF-8 encoding [7]. UTF-8 makes use of multiple bytes and represents complete collection of Unicode characters. This is achieved with the mechanisms of decoding and encoding as shown in Figure 5. Decoding translates text in files in a particular encoding like the Igbo text written with Igbo character sets into Unicode while encoding write Unicode to a file and convert it into an appropriate encoding [14]. A sample of an Igbo text is displayed in figure 6. The sources for the Igbo text documents collection for the work are as follows:

1. Igbo Radio - Online News reports in Igbo language.
2. Rex Publications - Igbo Catholic Weekly Bulletin.
3. Microsoft Igbo Language Software Localization Data
4. Igbo Novels

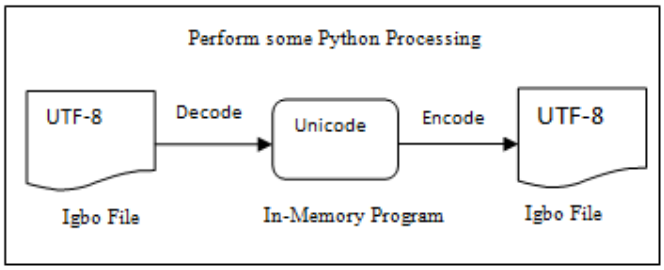


Figure 5. Igbo Text Unicode Decoding and Encoding

Kpaacharụ anya makana projektọ nkụziihe a achọghị okwu ntụghe, ndị ịchọghị ka ha hụ ga ahụ ihe-ngosi gi. Ọbụrụ na ịchọrọ iji projektọ nkụziihe a were rụọ ọrụ, pịkịnye "Jịkọọ". A na-akwụnye projektọ nkụziihe na kọmputa nkunaka iji mee ihe onyonyo. kọmputa nkunaka banyere na projektọ nkụziihe ọcha

Figure 6: Sample of Igbo Text Document

## 5.2 IGBO TEXT PRE-PROCESSING

Text pre-processing is an essential task and an important step in any text-based system. This module transformed unstructured input of Igbo text into a more understandable and structured format ready for further processing [15]. The text pre-processing task in this work covers text normalization, Igbo text tokenization and Igbo stop-words removal.

### 5.2.1 IGBO Text Normalization

In Normalization task, the Igbo textual document is transformed to a format that makes its contents consistent, convenient and full words for an efficient processing. *A*ll text cases are converted to lower cases. The diacritics and noisy data are removed. The noisy data is assumed to be data that are not in Igbo dataset and can be from:

- Numbers: Numbers can be cardinal numbers (single digits: 0-9 or a sequence of digits not starting with 0); signed numbers (contains a sign character (+ or -) following cardinal numbers immediately).
- Currency: This is mostly symbols used for currency e.g. £ (pound sign), € (Euro sign), ₦ (Naira sign), $ (dollar sign).
- Date and Time
- Other symbols like punctuation marks (:, ;, ?, !, ’ ), and special characters like <, >, /, @, “, !, *, =, ^, %, and others.

A list is created for these data and the normalization task process is done following the algorithm 1.

Algorithm 1**:** Algorithm for Igbo Text Normalization

Input: Igbo Text Document, Non-Igbo Standard Data/Character list

Output: Normalized Igbo Text

Procedure:

1. Transform all text cases to lower cases.
2. Remove diacritics (characters like ụ̄, ụ̀, and ụ́ contains diacritics called tone marks).
3. Remove non-Igbo standard data / character.
4. For every word in the Text Document:

    - If the word is a digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) or contains digits then the word is not useful, remove it.
    - If the word is a special character (:, ;, ?, !, ', (, ), {, }, +, &, [, ], <, >, /, @, ", !, *, =, ^, %, and others ) or contains special character, the word is non-Igbo, filter it out.
    - If the word is combined with hyphen like "nje-ozi", "na-aga", then remove hyphen and separate the words. For example, the following word "nje-ozi" will be "nje" and "ozi", two different words.
    - If the word contains apostrophe like n'elu, n'ụlọ akwụkwọ then remove the apostrophe and separate the words. For example "n'ụlọ akwụkwọ, after normalization will be three words "n", "ụlọ" and "akwụkwọ".

### 5.2.2 IGBO TEXT TOKENIZATION

Tokenization is the task of analyzing or separating text into a sequence of discrete tokens (words). The tokenization procedure used in the system is shown in algorithm 2.

Algorithm 2: Algorithm to tokenize the Igbo text
Input: Normalized Igbo text
Output: Tokenized Igbo Text

Procedure:

1. Create a TokenList.
2. Add to the TokenList any token found.
3. Separate characters or words between "-", if the string matches any of the following: "ga-", "aga-", "n'", "na-", "ana-", "ọga-", "ịga-", "ọna-", "ịna-". For instance, the following strings: "na–ese", "aga-eche", "na-eme" in a document will be separated into "na", "-", "ese", "aga", "-", "eche", "na", "-", and "eme" tokens.
4. Separate character or word(s) following n with apostrophe "n' ",. For instance, the following strings: "n'aka", "n'ụlọ egwu" in a document will be separated into "n", "aka", "n", "ụlọ" and "egwu" tokens.
5. Remove diacritics. This involves any non-zero length sequence of a–z, with grave accent ( ` ), or acute accent ( ´ ), for example, these words ìhè and ájá appearing in a given corpus will be taken as ihe and aja tokens, removing their diacritics.
6. Any string separated with a whitespace is a token.
7. Any single string that ends with comma (,) or colon (:) or semi-colon (;) or exclamation mark (!) or question mark (?) or dot (.), should be treated as a token.

Figure 6 shows the illustration of the result obtained by the Igbo Text Pre-processing System after performing text tokenization operation.

### 5.2.3 IGBO STOP-WORDS REMOVAL

Stop-words are language-specific functional words; the most frequently used words in a language that usually carry no information [12] [16]. There are no specific amount of stop-words which all Natural Language Processing (NLP) tools should have.

Most of the language stop-words are generally pronouns, prepositions, and conjunctions. This task removes the stopwords in Igbo text. Some of Igbo stop-words are shown in Figure 7.

ndị, nke, a, i, ị, o, ọ, na, bụ, m, mụ, ma, ha, ụnụ, ya, anyị, gị, niine, nile, ngị, ahụ, dum, niile, ga, ka, mana, maka, makana, tupu, e, kwa, nta, naanị, ugbua, olee, otu, abụọ, atọ, anọ, ise, isii, asaa, asatọ, iteghete, iri, anyi, ndị, a, n', g', ụfọdu, nari, puku, si, gara, gwa, ihi, dịka

Figure 7. Sample of Igbo Stop-words List

In the proposed system, a stop-word list is created and saved in a file named "stop-words" and is loaded to the system whenever the task is asked to perform. Any Igbo word with less than three character length is assumed not to carry useful information and is removed in this process. The removal of the stop words in the proposed system is done following the designed algorithm 3.

Algorithm 3: Algorithm to Remove Igbo Stop-Words
Input: Tokenized Igbo Text
Output: Stop-Word Free Text

Procedure:

1. Read the stop-word file.
2. Convert all loaded stop words to lower case.
3. Read each word in the created Token List.
4. For each word w ϵ Token List of the document
    - Check if w(Token List) is in Language stop-word list
    - Yes, remove w(Token List) from the Token List
    - Decrement tokens count
    - Move to the next w(Token List)
    - No, move to the next w(Token List)
    - Exit Iteration Loop
    - Do the next task in the pre-processing process
5. Remove any word with less than three (3) character length.

### 5.3 TEXT REPRESENTATION

Text representation involves the selection of appropriate features to represent a document [17]. The approach in which text is represented has a big effect in the performance of any text-based applications [18]. It is strongly influenced by the language of the text.

In Igbo language, compounding is a common type of word formation and many compound words exist. Compound words play high roles in the language. They can be referred as Igbo phrases that make sense only if considered as a whole. Majority of Igbo terms, key words or features are in phrasal structure. The semantic of a whole is not equal to the semantic of a part.

N-gram model is adopted to represent Igbo text because of the compound nature of the language [19]. The "N" spanned across 1 to 2, that is unigram and bigram. Unigram adopts the Bag-Of-Words (BOW) model and represents the Igbo text in single words. Bigram represents the Igbo text in sequence of two (2) words. The result of the two models on the feature selection is analyzed to find the n-gram model that gives the best relevant features on the feature selection of Igbo text documents.

### 5.4 FEATURE SELECTION

Feature selection is the task of choosing relevant features from a textual document to be used for a text-based task. This is put in place to reduce the dimension feature space of a system to improve its performance. It involves the identification of relevant features to be used in the system without affecting its accuracy [20]. This model will serve as a filter; muting out irrelevant, unneeded and redundant attributes / features from Igbo textual data to boost the performance of the system. Improving the feature selection will improve the system performance. The goal of this section (feature selection) is summarised in threefold:

1. Reducing the amount of features;
2. Focusing on the relevant features; and
3. Improving the quality of features used in the system process.

The Mean Term Frequency-Inverse Document Frequency (Mean TF-IDF) model is adopted for the feature selection in this work.

#### 5.4.1 MEAN TF-IDF

TF-IDF is a weighting filter method of feature selection, used to evaluate how essential a word is to text corpora. TF-IDF term weighting method is used on n-gram based term sequence to give suitable weights to the features generated to address the problem of high feature space dimensionality.

According to [7], TF-IDF of the features *t* in text document *d* is computed as follows:

$$\text{TF-IDF (t,d)} = \text{TF(t)} * \text{IDF (t)} \tag{1}$$

The normalized TF-IDF weight of the feature, t in a textual document d and is given by

$$W_{t,d} = TF_{t,d}(\log_{10}\frac{N}{DF_t}) = TF_{t,d} * IDF_t \tag{2}$$

The TF-IDF statistical weight model comprises of two parts:

1. The 1st part calculates the normalized Term Frequency (TF).

   TF(s) = (Number of times features s occurs in a document) / (Total number of features in the document).

   Term Frequency is defined as follows

$$TFi, j = \frac{ni, j}{\sum_n nk, j} \tag{3}$$

Where $n_{i,j}$ is the number of feature ($t_i$) occurences in document $d_j$; denominator is the sum of number of occurrences of all features in document $d_j$.

2. The 2nd part computes the IDF. It is calculated to measure the general importance of the feature / term.

   IDF(w) = Log (total text documents in the corpus / total of text documents where the specific feature occurs).

   The equation is as follows:

$$IDFi = \log\frac{|D|}{1+|\{d : ti \in d\}|} \tag{4}$$

In equation (4), 1 is added to the denominator (1+|{d:ti∈ d}) to avoid "division by zero" error if a term is not in the document.

The features with higher Mean TF.IDF values are selected. The selected features will then be used in further process in any intelligent text-based on the language.

Table 1. Mean TF-IDF Feature Selection Pattern

| **Terms** | **Text Documents** | | | | | **Mean(TF-IDF)** |
|---|---|---|---|---|---|---|
| | **D1** | **D2** | **D3** | **...** | **Dn** | |
| T1 | T1F1-IDF1 | T1F2-IDF2 | T1F3-IDF3 | **...** | T1Fn-IDFn | Mean(TF-IDF)1 |
| T2 | T2F1-IDF2 | T2F2-IDF2 | T2F3-IDF3 | **...** | T2Fn-IDFn | Mean (TF-IDF)2 |
| T3 | T3F1-IDF1 | T3F2-IDF2 | T3F3-IDF3 | **...** | T3Fn-IDFn | Mean (TF-IDF)3 |
| . | . | . | . | | . | . |
| . | . | . | . | | . | . |
| . | . | . | . | | . | . |
| Tn | TnF1-IDF1 | TnF2-IDF2 | TnF3-IDF3 | **...** | TnFn-IDFn | Mean (TF-IDF)n |

The sketch or pattern of Mean TF-IDF feature selection model used in the work is shown in table 1.

**Illustration of Mean TF-IDF**: The Mean TF-IDF model for feature selection employed in the proposed system is illustrated using two documents, D1 and D2.

**D1**: Kpaacharụ anya makana prọjektọ nkụziihe a achọghị okwu ntụghe, ndị ịchọghị ka ha hụ ga ahụ ihe-ngosi gi. Ọbụrụ na ịchọrọ iji prọjektọ nkụziihe a were rụọ ọrụ, pịkịnye "Jịkọọ". A na-akwụnye prọjektọ nkụziihe na kọmputa nkunaka iji mee ihe onyonyo. kọmputa nkunaka banyere na prọjektọ nkụziihe ọcha.

**D2:** Okwu ntụghe nke kọmputa nkunaka gị a dịkwazighị ịre ịjikọ na prọjektọ nkụziihe. ị ga a gbanwerịrị ya. kọmputa nkunaka na prọjektọ nkụziihe mara mma.

Table 2. Mean TF-IDF Feature Selection Illustration

| **Terms** | **Occurrence** | | **TF** | | **IDF** | | **TF-IDF** | | **Mean TF-IDF** |
|---|---|---|---|---|---|---|---|---|---|
| | **D1** | **D2** | **D1** | **D2** | **D1** | **D2** | **D1** | **D2** | |
| prọjektọ nkụziihe | 4 | 2 | 0.400 | 0.286 | 0.176 | 0.176 | 0.070 | 0.050 | 0.060 |
| kọmputa nkunaka | 2 | 2 | 0.200 | 0.286 | 0.176 | 0.176 | 0.035 | 0.050 | 0.043 |
| ihe onyonyo | 1 | 0 | 0.100 | 0.000 | 0.000 | 0.301 | 0.000 | 0.000 | 0.000 |
| okwu ntụghe | 1 | 1 | 0.100 | 0.143 | 0.176 | 0.176 | 0.018 | 0.025 | 0.021 |
| onyonyo komputa | 1 | 0 | 0.000 | 0.000 | 0.000 | 0.301 | 0.000 | 0.000 | 0.000 |
| nkunaka prọjektọ | 0 | 1 | 0.000 | 0.143 | 0.301 | 0.000 | 0.000 | 0.000 | 0.000 |
| anya prọjektọ | 1 | 0 | 0.100 | 0.000 | 0.000 | 0.301 | 0.000 | 0.000 | 0.000 |
| nkụziihe mara | 0 | 1 | 0.000 | 0.143 | 0.301 | 0.000 | 0.000 | 0.000 | 0.000 |

Using the Mean TF-IDF result shown in table 2, the mean TF-IDF average is 0.0155. The selection criterion is to select features with Mean TF-IDF greater than the Mean TF-IDF average (0.0155). The terms that will be selected as relevant features for further processing are "prọjektọ nkụziihe", "kọmputa nkunaka" and "okwu ntụghe" with Mean TF-IDF values 0.60, 0.043 and 0.021 respectively.

## 6. EXPERIMENT

This involves the practical method of putting into work all the theoretical design of the model. The Igbo feature selection model is designed with Object-Oriented methodology and implemented with Python programming language with tools from Natural Language Toolkit (NLTK). Many Igbo text documents were loaded to the system as shown in figure 8 and two documents (Igbo text2 and Igbo text6) were selected for the experiment. Figure 9 and figure 10 display the result of feature selection on Igbo unigram text and bigram text respectively.

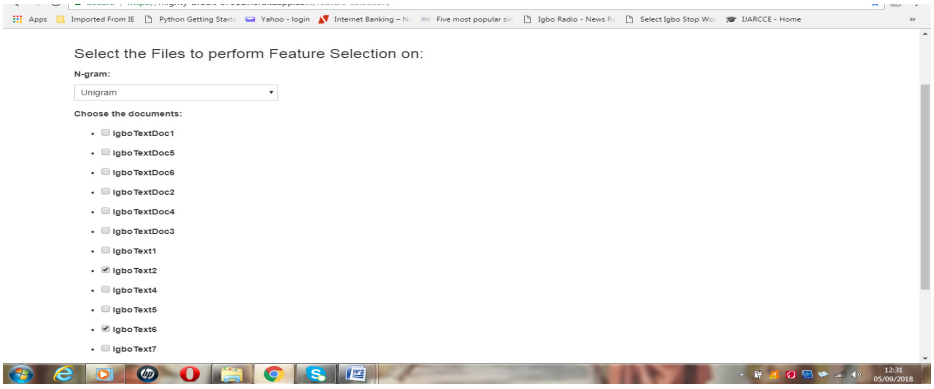


Figure 8. Screen Display of Igbo Feature Selection System

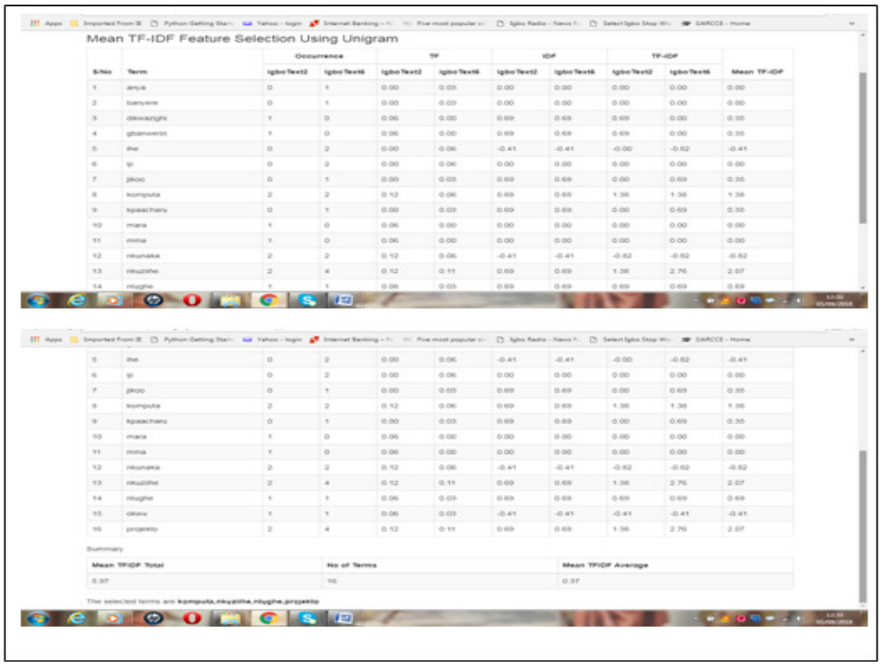


Figure 9. Feature Selection Result on Igbo Unigram Represented Text


Figure 10. Feature Selection Result on Igbo Bigram Represented Text

## 7. RESULT DISCUSSIONS

Figure 8 is a screen display of the Feature Selection System to perform the selection of relevant features from Igbo textual documents for further text processing. Many Igbo textual documents are loaded to the system. For the experiment and result discussions, the feature selection was performed on two documents (Igbo text2 and Igbo text6). The threshold values for the selection of the most relevant features are set to greater than Mean TF-IDF average.

Figure 9 shows result of the mean TF-IDF feature selection on unigram represented Igbo text. The selection criterion is to select features with Mean TF-IDF greater than the Mean TF-IDF average. As shown in the figure 9 result, the mean TF-IDF average of Igbo Text2 and Igbo Text6 in Unigram text is 0.37. Only 4 key features are selected, the most relevant features. These are “

Kọmputa – Computer (Mean TF-IDF is 1.38)", "Nkụziihe – Teaching (Mean TF-IDF is 2.07)", "Ntụghe – Opening (Mean TF-IDF is 0.69)", and "Prọjektọ- Projector (Mean TF-IDF is 2.07)".
Figure 10 shows the display of result obtained at the feature selection module when bigram representation of the selected texts is chosen. The result shows that three (3) features with Mean TF-IDF above the Mean TF-IDF average (0.53) of Igbo Text2 and Igbo Text6 in bigram representation are selected. The features are "kọmputa nkunaka – laptop (Mean TF-IDF is 1.36)", "okwu ntụghe – password (Mean TF-IDF is 1.36)", "prọjektọ nkụziihe – Teaching Projector (Mean TF-IDF is 2.07)". The model extracts bigrams that accurately describe relevant compound words in Igbo language and the context in which the words are used.

## 8. CONCLUSION

The development in Information Technology (IT) has encouraged the use of Igbo Language in text creation, online news reporting, online searching and articles publications. As the information stored in text format of this language is increasing, an efficient model that selects most relevant features from Igbo textual documents has been designed and implemented. This is to improve the performance and accuracy of any intelligent text-based system on the language when adopted. The system corpora is richly represented using n-gram model (unigram and bigram) before feature selection is performed on it to address the issue of compounding which plays high role in the language. The unigram selected features do not dully represent the actual context in which the words are used in the documents but the bigrams features accurately described the context the words are used as well as the compound words in the language. This indicates that feature selection on bigram represented text is recommended for text-based system in Igbo language.
The model is highly recommended for any Igbo Intelligent text-based system because it will certainly improve the system accuracy and performance; reduce feature dimension space and system processing time as well as the computational cost. It will also improve data and model understanding in the language.

## ACKNOWLEDGEMENTS

The authors would like to appreciate the reviewers of this paper, for their useful comments and contributions which added more to the quality of the work.

## REFERENCES

[1] Pradnya Kumbhar and Manisha Mali (2016). A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. International Journal of Science and Research (IJSR). Volume 5 Issue 5, pp 1267 - 1275.

[2] Veerabhadrappa and Lalitha Rangarajan (2010). Multi-Level Dimensionality Reduction Methods using Feature Selection and Feature Extraction. International Journal of Artificial Intelligence & Applications (IJAIA), Volume 1, No.4, pp 54 - 68

[3] Aisha Adel, Nazlia Omar and Adel Al-Shabi (2014). A comparative Study of Combined Feature Selection Methods for Arabic Text Classification. Journal of Computer Science. Vol. 10, No.11, pp 2232 – 2239.

[4] Noura alnuaimi, Mohammad M Masud and Farhan Mohammed (2015). Examining The Effect Of Feature Selection On Improving Patient Deterioration Prediction. International Journal of Data Mining & Knowledge Management Process (IJDKP) Volume5, No.6, pp 13 -33

[5] Ladha L. and Deepa T. (2011). Feature Selection Methods and Algorithms. International Journal on Computer Science and Engineering (IJCSE). Volume 3 No.5, pp 1787 – 1797.

[6] Sunita Beniwal and Jitender Arora (2012). Classification and Feature Selection Techniques in Data Mining. International Journal of Engineering Research & Technology (IJERT). Vol. 1 Issue 6, pp 1-6.

[7] Ifeanyi-Reuben, N.J., Ugwu, C. and Adegbola, T. (2017). Analysis and representation of Igbo text document for a text-based system. International Journal of Data Mining Techniques and Applications (IJDMTA). Vol. 6, No. 1, pp 26-32.

[8] Bilal Hawashin, Ayman M. Mansour and Shadi Aljawarneh (2013). An Efficient Feature Selection Method for Arabic Text Classification. International Journal of Computer Applications (0975 – 8887). Vol. 83, No.17, pp 1 -6.

[9] Ghazi Raho, Ghassan Kanaan, Riyad Al-Shalabi and Asma'aNassar (2015). Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study. International Journal of Advanced Computer Science and Applications (IJACSA). Vol. 6, No. 2, pp 192 – 195.

[10] Rehab Duwairi, Mohammad Nayef Al-Refai and Natheer Khasawneh (2009). Feature Reduction Techniques for Arabic Text Categorization. Journal of the American Society for Information Science and Technology. Vol. 60, No. 11, pp 2347–2352.

[11] Bird, S., Klein, E. and Loper, E. (2009). Natural language processing with Python. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[12] Mohammad Aizat Bin Basir and Faudziah Binti Ahmad (2017). New Feature Selection Model-Based Ensemble Rule Classifiers Method For Dataset Classification. International Journal of Artificial Intelligence and Applications (IJAIA), Vol.8, No.2, pp 37 – 43

[13] Sabu M.K (2015). A Novel Hybrid Feature Selection Approach for the Prediction of Learning Disabilities In School-aged Children. International Journal of Artificial Intelligence & Applications (IJAIA) Vol. 6, No. 2, pp 67 -80

[14] Arjun S. N., Ananthu P. K., Naveen C. and Balasubramani R. (2016). Survey on pre-processing techniques for text mining. International Journal of Engineering and Computer Science. Vol. 5, No. 6, pp 16875-16879.

[15] Harmain M., H. El-Khatib and A. Lakas, (2004). Arabic Text Mining. College of Information Technology United Arab Emirates University. Al Ain, United Arab Emirates. IADIS International Conference Applied Computing 2004, Issue 2, pp 33 -38.

[16] Shen, D., Sun, J., Yang, Q. and Chen, Z. (2006). Text classification improved through multi-gram models," In Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management (ACM CIKM 06), Arlington, USA. Pp 672-681.

[17] Raed Al-Khurayji and Ahmed Sameh (2017). An Effective Arabic Text Classification Approach Based on Kernel Naive Bayes Classifier. International Journal of Artificial Intelligence and Applications (IJAIA), Vol.8, No.6, pp 1-10

[18] David, D.L. (1990). Representation quality in text classification: An Introduction and Experiment. Selected papers from the AAAI Spring Symposium on text-based Intelligent Systems. Technical Report from General Electric Research & Development, Schenectady, NY, 1230.

[19] Ifeanyi-Reuben, N.J., Ugwu, C. and Nwachukwu, E.O. (2017). Comparative Analysis of N-gram Text Representation on Igbo Text Document Similarity. International Journal of Applied Information Systems (IJAIS), Vol.12, No. 9, pp 1-7.

[20] Divya P. and Nanda K. G. S. (2015). Study on feature selection methods for text mining. International Journal of Advanced Research Trends in Engineering and Technology (IJARTET). Vol. 2, No. 1, pp 11- 19.

## AUTHORS

**Ifeanyi-Reuben Nkechi J.** has a doctorate degree in Computer Science from the University of Port-Harcourt Nigeria. She obtained her M.Sc. and B.Sc. in Computer Science from the University of Ibadan Nigeria and University of Calabar Nigeria respectively. She is a lecturer at the Department of Computer Science, Rhema University Nigeria. She is a member of Computer Professionals (Registration Council) of Nigeria (CPN), Nigeria Computer Society (NCS) and Nigeria Women in Information Technology (NIWIIT). Her research interests include Database, Data mining, Text mining, Information Retrieval and Natural Language Processing.

**Benson-Emenike Mercy E.** has a doctorate degree and Masters degree in Computer Science from University of Port Harcourt, Nigeria. She obtained her Bachelor of Technology degree [B.Tech] from Federal University of Technology, Minna, Niger state. She is a lecturer in the Department of computer Science, Abia State Polytechnic and an adjunct lecturer in Computer science Department, Rhema University Nigeria and National Open University of Nigeria [NOUN]. She is a member of Computer Professionals (Registration Council) of Nigeria (CPN) and Nigeria Computer Society (NCS). Her research interests include Artificial Intelligence, Biometrics, Operating System, and Information Technology.