Semantic Representation of Igbo Text Using Knowledge Graph*

Nkechi J.Ifeanyi-Reuben[1,†], Patience Usoro Usip[2,*,†]

[1]Computer Science Department, Nnamdi Azikiwe University Awka, Nigeria)
[2]Computer Science Department, University of Uyo, Uyo, Nigeria

Abstract

With the fast growth of Artificial Intelligence and its application in different areas of Natural Language Processing, semantic representation contributes immensely to smoothing the progress of different automated language processing applications. Semantic representation returns the meaning of the text as it may be understood by humans. Although semantic representation is very useful for several applications, no semantic model is proposed for the Igbo language. The usage of Igbo language in the text-based applications such as text mining, information retrieval, natural language processing is at the increase. Igbo language uses compounding in its word formation and word ordering play high role in the language. The uncertainty in dealing with these compound words has made the representation of Igbo text very difficult. There is need to for smart data representation model in the said language to enhance efficiency and effectiveness in its text-based application. This paper presents the analysis of a language classification, considering Igbo language, considering its compounding nature and describes a smart model for text representation using a Knowledge Graph. The model will create a smart data repository the real-world usage of text and tangled its context relationship. The proposed Igbo Knowledge Graph (IKG) text representation model was used in Igbo text classification system. The performance of the Igbo text classification system is measured by computing the precision, recall and F1-measure of the result obtained on bigram, semantic-based and unigram represented textual documents. The Igbo text classification on semantic-based represented text has highest degree of exactness (precision). This shows that the classification on semantic-based Igbo represented text outperforms bigram and unigram represented texts. Semantic-based text representation model using knowledge graph is highly recommended for any Igbo text-based system. It enables automated reasoning as well addresses the challenges incurred as a result of Igbo compounding, word ordering and collocations language peculiarities.

Keywords

Igbo Language, Text Representation, Text Classification, Ontology, Knowledge Graph, Artificial Intelligence, Compound Word, Semantics

IWMSW-2022: International Workshop on Multilingual Semantic Web, Co-located with the KGSWC-2022, November 21–23, 2022, Madrid, Spain

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

**Corresponding author.

†These authors contributed equally.

nj.ifeanyi-reuben@unizik.edu.ng (N. J.Ifeanyi-Reuben); patienceusip@uniuyo.edu.ng (P. U. Usip)

0000-0002-6516-5194 (P. U. Usip)

1. Introduction

Text representation is the selection of appropriate features to represent document [1]. The approach in which text is represented has a big effect in the performance of any text-based applications [2]. It is powerfully controlled by the language of the text. The spread of Information Technology (IT) in real life activities has assisted in inculcating Igbo language in text-based application such as text creation, web creation, text mining, information retrieval and natural language processing. This research improved the existing research of [3]. on the analysis and representation of Igbo text for a text-based system by incorporating the semantic representation of the text in order to create detailed notations of the text that accurately conveys its meaning. Semantic representation of the textual document is very rich and is adopted in many applications of Natural Language Processing (NLP) such as machine translation, information retrieval, question answering, text classification, sentiment analysis, text summarisation and text extraction. It reflects the meaning of the text as it may be understood by humans. Thus, it contributes to facilitating various automated language processing applications. The research on [3, 4, 5, 6] emphasized that the semantic representation of Arabic text can facilitate several natural language processing applications such as text summarization and textual entailment. Semantic representation can be achieved using Knowledge Graph (KG). Semantic representation reflects the meaning of the text as it may be understood by humans. Thus, it contributes to facilitating various automated language processing applications. Semantic representation can be achieved using Knowledge Graph (KG).

Knowledge Graph (KG) is a way to represent and organize the data in a more efficient and easy way to modify, use, and understand [7] It is also referred to as a collection of interlinked description of concepts, entities, relationships and events via linking and semantic metadata, providing a framework for data integration, unification, analytics and sharing.

With the widespread growth of Igbo data on the Web, the need for efficient methods to get and arrange valuable information from these big noisy data is increased. This research presents an Igbo Knowledge Graph (IKG) for representing data created with Igbo language for better performance for any Igbo text-based applications.

This Igbo smart representation will be useful for many purposes such as question answering, summarization and information retrieval.

The model chosen by the researchers will also help to discover unidentified facts and concealed knowledge that may exist in the lexical, semantic or relations in Igbo text corpus.

1.1. Language classification

A language is a method of communication between individuals who share common code, in form of symbols [8]. In linguistics, there are two kinds of language classification: genetic (or genealogical) and typological.

Genetic, also known as genealogical language is a type that group languages into families based to their degree of diachronic relatedness. Examples of genealogic language group are German, English, Dutch, Swedish, Norwegian, Danish, Irish, Welsh, Breton, etc.

Typological classification groups languages into types according to their structural characteristics. These structural characteristics can be phonological typology, morphological typology or

syntactic typology. Typological languages form words by agglutination. Examples are Igbo, Turkish, Finnish, Japanese, etc. [9]

Igbo Language The Igbo language is one of the agglutinative languages, a language that form words through the combination of smaller morphemes to get compound words. It is one of the three major languages in Nigeria. It is largely spoken by the people in the eastern part of Nigeria. Igbo language has many dialects. The standard Igbo is used formally and is adopted for this research. The current Igbo orthography [8] is based on the Standard Igbo. Orthography is a way of writing sentence or constructing 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, nw, ny, ṅ, o, ọ, p, r, s, sh, t, u, ụ, v, w, y, z). Igbo language has a large number of compound words. A compound word is a word that has more than one root, and can be made from combination of either nouns, pronouns or adjectives.

Ifeanyi-Reuben et al. [8] studied the Igbo compound words and categorized them as follows:

i. Nominal (NN) Compound Word: A nominal compound word is formed by the combination of two or more nouns. The nominal compound words are written separately not minding the semantic status of the nouns in Igbo. Example of Igbo nominal compound words are: nwa akwụkwọ - student; onye nkuzi – teacher; ama egwuregwu – stadium; ụlọ ọgwụ - hospital; ụlọ akwụkwọ - school.

ii. Agentive Compound Words: In agentive compound word, one or more nouns express the meaning of the agent, doer of the action. The Igbo agentive compound words are written separately irrespective of the translations in English. They can also be referred to as VN (Verb Noun) compound words. Example: oje ozi – messenger; oti ịgba - drummer.

iii. Igbo Duplicated Compound Word: Igbo duplicated compound words are formed by the repetition of the exact word two or more times to show a variety of meaning. For example: ọsọ ọsọ - quickly; mmiri mmiri – watery; ọbara ọbara – reddish.

iv. Igbo Coordinate Compound Words: This compound word is formed by the combination of two or words joined by the Igbo conjunction “na” meaning “and” in English. All the Igbo compound words of this category is written separately. Example: Ezi na ụlọ - family; okwu na ụka – quarrel.

v. Igbo Proper Compound Words: This category of Igbo compound words includes personal names, place names, and club names. All words in this category are wriiten together not minding how long they may be. Example: Uchechukwu; Ngozichukwuka; Ifeanyichukwu.

vi. Igbo Derived Compound Words: The derived Igbo compound words are words derived from verbs or phrases. The roots of the derived Igbo compound words are written together. Example: Dinweụlọ - landlord.

Igbo, being an agglutinative language, has a huge number of compounds words and can be referred to as a language of compound words. The proposed research of Igbo Knowledge Graph representation will consider this peculiarity to get a good result.

2. Related Works

Ifeanyi-Reuben et al. [chidiebere2020analysis] presents the analysis of Igbo language text document and describes its representation with the Word-based N-gram model. The result shows that Bigram and Trigram n-gram text representation models perform better than unigram model.
Wael and Arafat [3] proposed a graph-based semantic representation model for Arabic text. The proposed model aims to extract the semantic relations between Arabic words. The results proved that the proposed graph-based model is able to enhance the performance of the textual entailment recognition task in comparison to other baseline models.
Zhang, Yoshida and Tang [10] studied and compared the performance of adopting TF*IDF, LSI (Latent Semantic Indexing) together with multiple words for text representation. They used Chinese and English corpora to assess the three techniques in information retrieval and text categorization. Their result showed that LSI produced greatest performance in retrieving English documents and also produced best performance for Chinese text categorization.
Chih-Fong [11] improved and applied Bag of Word (BOW) for image annotation. An image annotation is used to allocate keywords to images automatically and the images are represented using characteristics such as color, texture and shape. This is applied in Content-Based Image Retrieval System (CBIRS) and the retrieval of the image is based on indexed image features.
Usip and Ntekop [12] posited that ontology is a necessary technology tool for easy and intelligent reasoning with knowledge. Being the underlying schema for every knowledge graph, this study will improve the existing work of Ifeanyi-Reuben et al. [8] by adding intelligence to the work using Knowledge Graph. Ontology-driven applications for multilinguality was described by Usip and Ekpenyong [13].
Etaiwi and Awajan [14] proposed SemG-TS, a novel semantic graph embedding-based abstractive text summarization model for the Arabic language which employed a deep neural network to generate abstractive summary. The result obtained shows SemG-TS model outperforms the popular baseline word embedding technique, word2vec.

3. Methodology

The bulk of concerns for any text-based system are attributed to text representation considering the peculiarities of the natural language involved. In this section, we propose an efficient and effective model to represent Igbo text to be adopted by any text-based system. This is a process of transforming unstructured Igbo textual document into a form proper for automatic processing. This is a vital step in text processing because it affects the general performance of the system. The proposed approach for the Igbo text representation process is shown in Figure 1.

3.1. Text Preprocessing

Text preprocessing involves tasks that are performed on text to convert the original natural language text to a structure ready for processing. It performs very important functions in

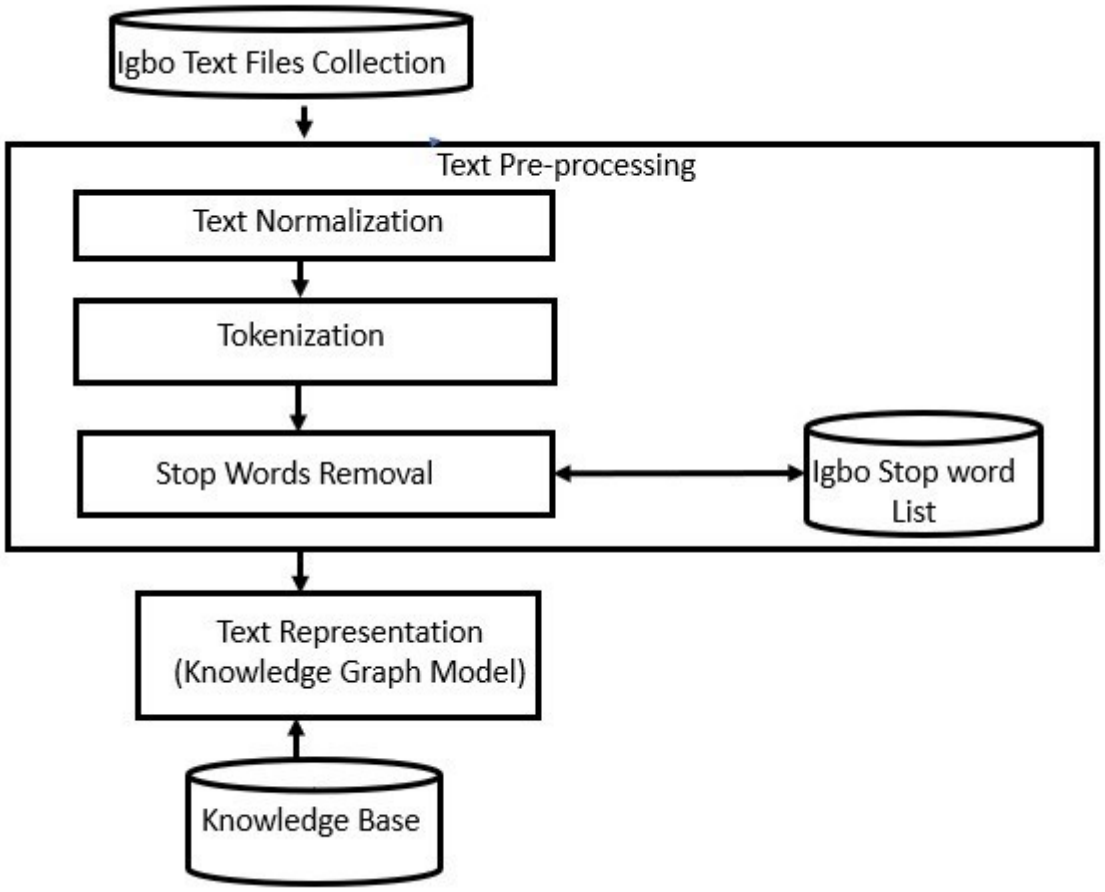

Figure 1: Igbo Text Representation Process.

different text-based system. The tasks are Igbo text normalization, Igbo text tokenization and Igbo text Stop words Removal.

Igbo Text Normalization: In Normalization process, we transformed the Igbo textual document to a format to make its contents consistent, convenient and full words for an efficient processing. We transformed all text cases to lower case and also removed diacritics and noisy data. The noisy data is assumed to be data that are not in Igbo dataset. Text Tokenization: Tokenization is the task of analyzing or separating text into a sequence of discrete tokens (words).

Igbo Stop-words Removal: Stop-words are language-specific functional words; the most frequently used words in a language that usually carry no information. There are no specific number 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 stopwords is shown in Figure 2.

In the proposed system, a stop-word list will be created and saved in a file named “stop-words” and is loaded to the system whenever the task is asked to perform.

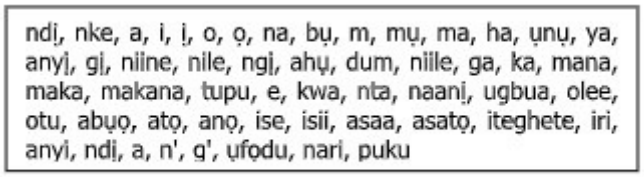

Figure 2: Sample of Igbo Stop-words list.

3.2. Knowledge Graph Text Representation

Knowledge graphs combine characteristics of several data management paradigms:
• Database: The data can be explored via structured queries.
• Graph: Data can be analyzed as any other network data structure.
• Knowledge base: The model will bear formal semantics, which can be used to interpret the data and infer new facts.
Igbo Knowledge graphs will provide good framework for Igbo data integration, unification, linking and reuse.

4. Sample Igbo Text and the Corresponding Proposed Knowledge Graph

Given the examples of Igbo compound words in table 1, it is observed that the actual meaning of the semantic correctness of Igbo compound words is not the same when compared with their roots and meaning after decomposition to Igbo single words. Hence, the need for the compound word categorization.

Following the categorization of the Igbo compound words, a knowledge graph representation of Igbo words and the various categories is given in Figure 3.

The underlying ontology used as the schema for the knowledge graph has the domain knowledge which includes the bilingual corpora of Igbo single words and their English meaning, the n-gram modeling feature and resulting Igbo compound words classified based on the compound word categorization.

From the knowledge graph, the relationship among the various Igbo compound word, single Igbo word and their English word meaning can be determined and used in an effort towards the construction of a semantically correct bilingual Igbo - English Language dictionary consisting of both single and compound Igbo words. With the knowledge graph, missing links between Igbo compound and single words can be detemined at a glance for proper restructuring and fixture to produce a semantically correct Igbo word

Table 1
Igbo Compound Words [8]

Igbo Compound Words	Meaning	Roots and meaning	Compound Word Category
Onye nkuzi	Teacher	Onye - Person Nkuzi – Teach	Nominal
Ezi na ụlọ	Family	Ezi – surrounding Na – and ụlọ - family	Coordinate
Ojiiegoachọego	businessman	Ojiiego – use money achọego – find money	Derived
ụgbọ ala	Car, motor	ụgbọ - vessel ala - land (road)	Nominal
Egbe igwe	Thunder	Egbe – gun Igwe – sky	Nominal
Iri abụọ	Twenty	Iri – ten Abụọ - two	Nominal
Ode akwụkwọ	Secretary	Ode – Write Akwụkwọ - book	Agentive
Eberechukwu	God's mercy	Ebere – mercy Chukwu - God	Proper
Mmiri mmiri	Watery	Mmiri -water	Duplicate
ọcha ọcha	Whitish	ọcha – white	Duplicate
Onye nchekwa	Administrator	Onye – person Nchekwa – protect	Nominal
Kọmputa Nkunaka	Laptop	Kọmputa – Computer Nkunaka – Handcarry	Nominal
Ọkpụ ụzụ	Blacksmith	ọkpụ - mold ụzụ - clay	Agentive
Nche anwụ	Umbrella	Nche – protect Anwụ - sun	Agentive
Onyonyo kọmputa	Monitor	Onyonyo – screen Kọmputa- computer	Nominal
Okwu ntughe	Password	Okwu – speech Ntughe - opening	Nominal

Figure 4 is a designed Igbo knowledge graph model showing all the due processes employed to represent Igbo textual document based on its semantic (reasoning) using knowledge graph.

5. System Performance Evaluation

The system performance is evaluated by computing the precision, F1-measure and Recall. Precision is defined as the quotient of total TPs and sum of total TPs and FPs. Precision point is known to as a point of correctness.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Recall of the classification system is described as the quotient of total TPs and sum of total

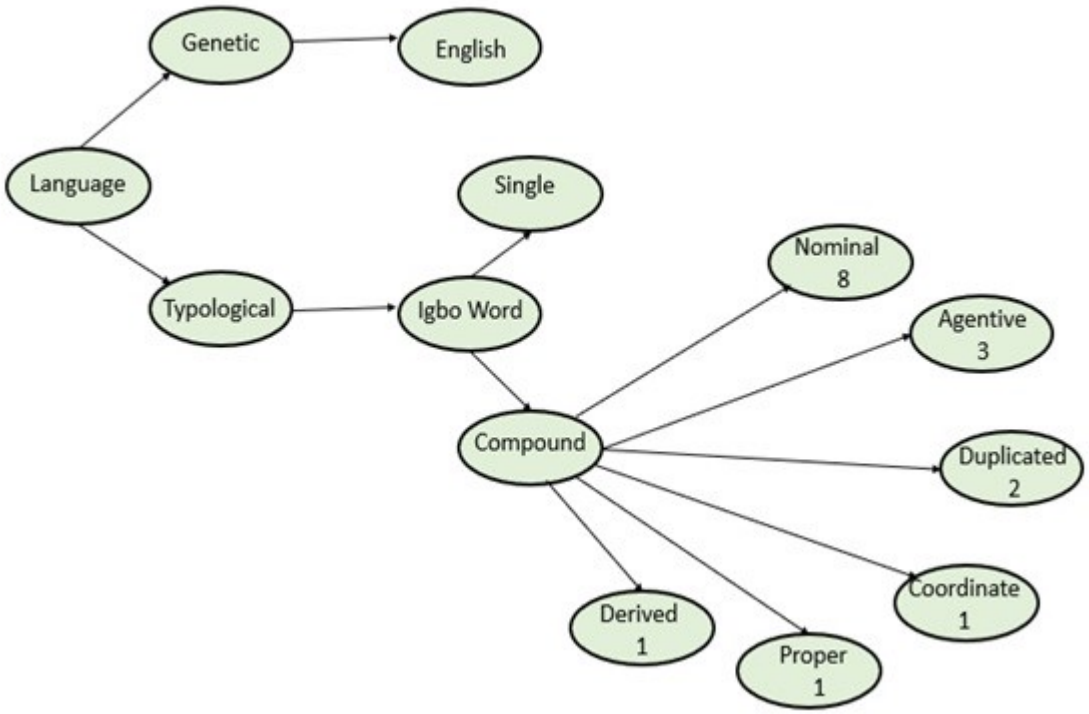

Figure 3: Knowledge Graph Representation of Sample Igbo Compound.

TPs and total FNs. Recall level measures completeness.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

F1-Measure is single function that joins recall and precision points. When the F1-measure is high, it means that the overall text classification system is high.

$$\text{F1-Measure} = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{3}$$

$$= \frac{2TP}{(2TP + FP + FN)} \tag{4}$$

In summary, computation of precision, recall and f1-measure required four input parameters: TP, FP, TN and FN.

i. TP - total of text documents accurately allotted to document class.

ii. FP - total of text documents wrongly allotted to document class.

iii. FN - total of text documents wrongly rejected from document class.

iv. TN - total of text documents correctly rejected from document class.

These parameters are input to the evaluator. They are obtained from the classification result.

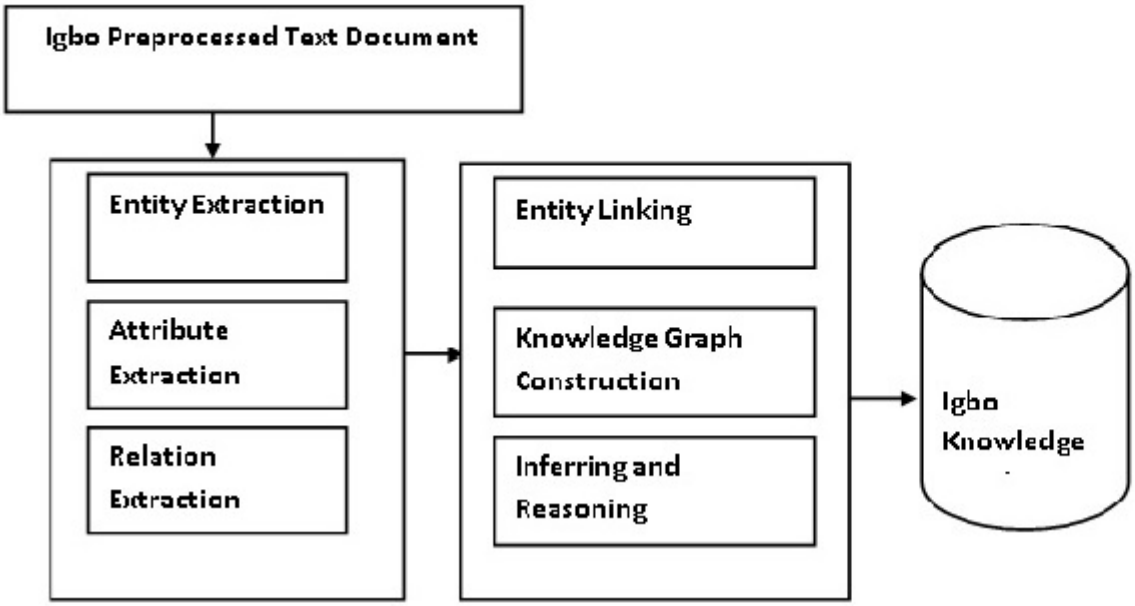

Figure 4: Igbo Knowledge Graph Model

6. Experiments

This involves the practical method of putting into work all the theoretical design of the proposed model. The semantic representation of Igbo text using knowledge graph is implemented on Igbo text classification system with Python and tools from Natural Language Toolkit (NLTK).

Table 2
Summary of Bigram, Semantic-Based and Unigram Text Classification result

Text Document	Bigram	Semantic-Based	Unigram
IgboText1	Administration	Administration	Administration
IgboText10	Religion	Religion	Religion
IgboText11	Computer	Computer	Administration
IgboText2	Religion	Computer	Religion
IgboText4	Administration	Administration	Administration
IgboText5	Religion	Administration	Administration
IgboText6	Computer	Computer	Administration
IgboText7	Administration	Administration	Administration
IgboText8	Religion	Religion	Religion
IgboText9	Religion	Religion	Religion

Table 3
Performance Measure Result

Category	TP	NP	FP	FN	Precision	Recall	F1 Measure
Bigram	8	0	2	0	0.80	1.00	0.89
Semantic-Based	9	0	1	0	0.90	1.00	0.95
Unigram	7	0	3	0	0.70	1.00	0.82

7. Result Analysis

Figure 5 displays the Text classification module of the system used to test the effectiveness of the proposed model. The result obtained in text classification on Igbo text represented semantically using knowledge graph is compared with the results obtained in unigram and bigram text representation models.

Table 3 and Figure 6 show the classification performance measure result and chart respectively. The result shows that the recall, precision and F1 for bigram Igbo represented text are 1.00, .80 and .89 respectively. The recall, precision and F1 for semantic-based Igbo text are 1.00, .90 and .95 respectively. The recall, precision and F1 for unigram Igbo represented text are 1.00, .62 and .82 respectively. Recall evaluates the degree of completeness. The result shows Igbo text classification on the text represented with the three models (bigram, semantic-based and unigram) has the equal level of recall (completeness). This means all the text documents that were given to the classifier, were given a label name. Precision measures the degree of exactness. The classification with semantic-based has highest degree (0.90) of exactness (precision).

Table 2 gives the summary of classification result obtained on Bigram, Semantic-Based and Unigram text representation. A total of 10 testing documents are used for the experiment. In bigram, eight documents are correctly assigned a class label while two are incorrectly assigned a class label. In semantic-based text representation using knowledge graph, 9 documents are correctly assigned a class label while one is incorrectly assigned. In unigram, 7 documents are correctly assigned a class label while 3 are incorrectly assigned a class label.

8. Conclusion

An improved intelligent approach for representing Igbo text document using Knowledge Graph model considering the agglutinative nature of Igbo language is proposed. This is to solve the issues of collocations, compounding, and word ordering that plays major roles in the language, thereby making the representation semantic-enriched. The model is implemented and evaluated using Igbo text classification system.

The performance was measured by computing the classification accuracy of Bigram, Semantic-Based and Unigram represented text. The result showed that the classification performed on Semantic-based represented text has higher performance than Bigram and unigram represented texts. It has shown that a high quality text representation model certainly boost performance of NLP tasks.

The model will be of high commercial potential value and will be useful in any text based intelligent system on the language. It will also motivate other researchers to develop interest in doing more research on Igbo language processing to the benefit of people and society.

9. Acknowledgments

The authors wish to express gratitude the unknown reviewers of this work for their useful comments and contributions that assisted in enhancing the worth of this paper.

References

[1] D. Shen, J.-T. Sun, Q. Yang, Z. Chen, Text classification improved through multigram models, in: Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 672–681.

[2] D. D. Lewis, Representation quality in text classification: An introduction and experiment, in: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990, 1990.

[3] W. Etaiwi, A. Awajan, Graph-based arabic text semantic representation, Information Processing & Management 57 (2020) 102183.

[4] S. Tiwari, P. Siarry, S. Mehta, M. Jabbar, Tools, Languages, Methodologies for Representing Semantics on the Web of Things, John Wiley & Sons, 2022.

[5] R. Panchal, P. Swaminarayan, S. Tiwari, F. Ortiz-Rodriguez, Aishe-onto: a semantic model for public higher education universities, in: DG. O2021: The 22nd Annual International Conference on Digital Government Research, 2021, pp. 545–547.

[6] F. Ortiz-Rodriguez, S. Tiwari, R. Panchal, J. M. Medina-Quintero, R. Barrera, Mexin: Multidialectal ontology supporting nlp approach to improve government electronic communication with the mexican ethnic groups, in: DG. O 2022: The 23rd Annual International Conference on Digital Government Research, 2022, pp. 461–463.

[7] I. A. Ahmed, F. N. AL-Aswadi, K. M. Noaman, et al., Arabic knowledge graph construction: a close look in the present and into the future, Journal of King Saud University-Computer and Information Sciences (2022).

[8] U. Chidiebere, A. Tunde, et al., Analysis and representation of igbo text document for a text-based system, arXiv preprint arXiv:2009.06376 (2020).
[9] M. Onukawa, The writing of standard igbo in okereke oo (ed.) readings in citizenship education, Okigwe: Wythem Publishers (2001).
[10] W. Zhang, T. Yoshida, X. Tang, Text classification based on multi-word with support vector machine, Knowledge-Based Systems 21 (2008) 879–886.
[11] C.-F. Tsai, Bag-of-words representation in image annotation: A review, International Scholarly Research Notices 2012 (2012).
[12] P. U. Usip, M. Ntekop, The use of ontologies as efficient and intelligent knowledge management tool, in: 2016 Future Technologies Conference (FTC), IEEE, 2016, pp. 626–631.
[13] P. U. Usip, M. E. Ekpenyong, Towards ontology-driven application for multilingual speech language therapy, in: Human Language Technologies for Under-Resourced African Languages, Springer, 2018, pp. 85–101.
[14] W. Etaiwi, A. Awajan, Semg-ts: Abstractive arabic text summarization using semantic graph embedding, Mathematics 10 (2022) 3225.

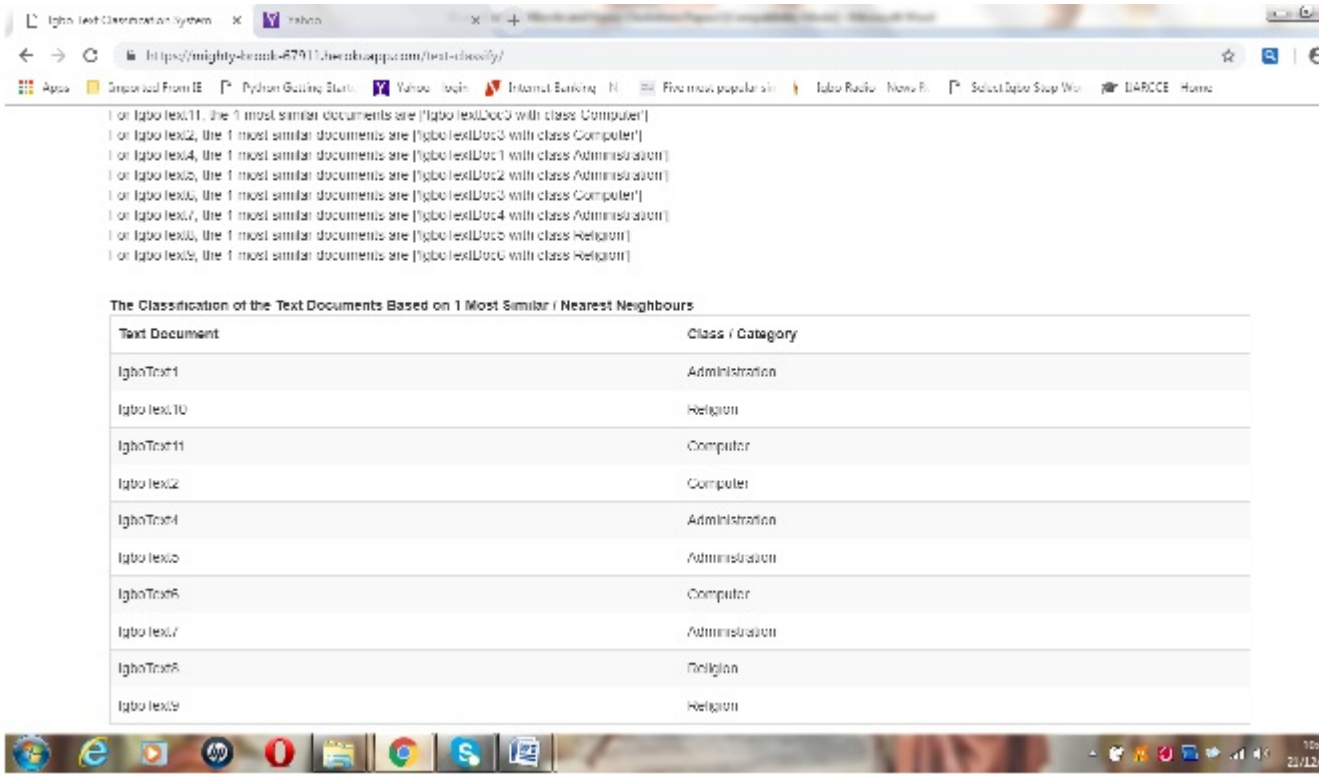

Figure 5: Igbo Text Classification System Result

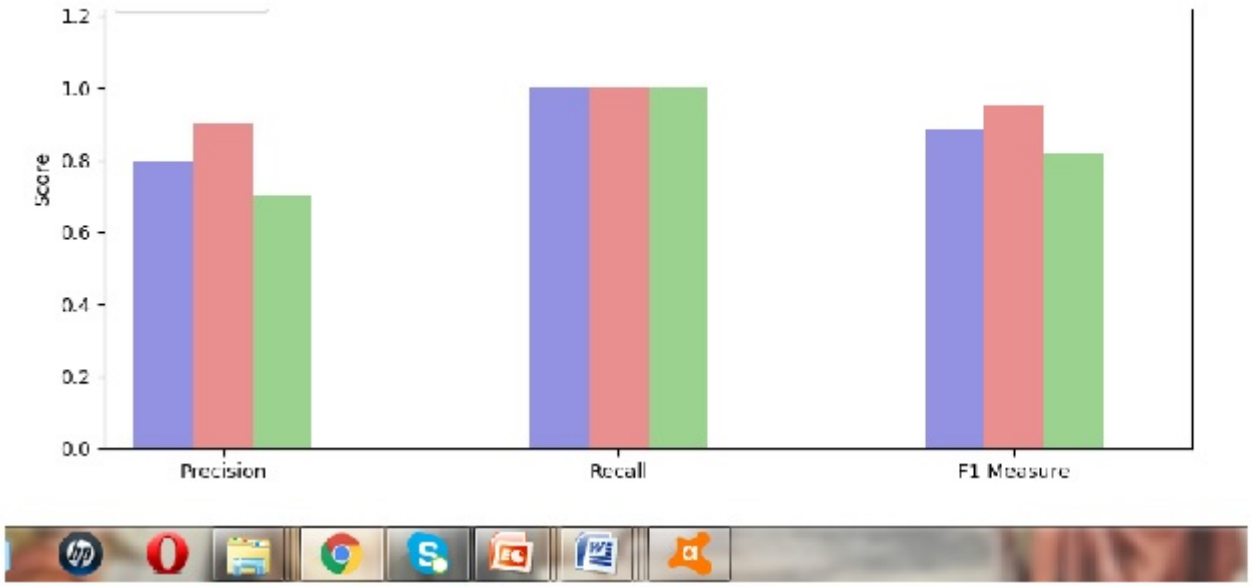

Figure 6: Igbo Text Classification System Performance Measure Result Chart