World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

Design and implementation of an automated web-based Igbo text analyzer using natural language processing (NLP) tools

Prince C. Azubuike * and Innocent I Umeh [2]

Department of Computer Science, Nnamdi Azikiwe University Awka, Anambra State, Nigeria.

World Journal of Advanced Research and Reviews, 2024, 23(03), 1036–1045

Publication history: Received on 03 June 2024; revised on 04 September 2024; accepted on 06 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2691

Abstract

Presently in the world, the Igbo language is one of the less-resourced languages because there are not many developed and easy-to-find digital resources for it. Digital resources such as Igbo text corpora, Igbo electronic dictionaries, Igbo morphological analyzers, and Igbo thesauri, which can analyze Igbo text documents, are very limited. This work aims to design and develop an automated Igbo text analyzer using Natural Language Processing (NLP) tools. The development of this web-based Igbo text analyzer involves the analysis of the lexical and grammatical characteristics of the Igbo language that aided the identification of the basic principles governing word change (inflection) in the Igbo language. The object-oriented hypermedia design methodology (OOHDM) was applied to segment the work into stages of conceptual design, navigational design, abstract interface design, and implementation. The system was implemented using ReactJS for the frontend and the Python Flask framework for the backend. Furthermore, SQLite and SQLiteStudio were used as the database and database management tools for the system. The Natural Language Toolkit (NLTK) was used for text document analysis to enable users to observe the frequency and statistical analysis of the Igbo text document, as well as the Part of Speech (POS) tags associated with the words of the language. The development of this Igbo text analyzer (IgboNLP web application) is a great step towards achieving the objectives of Basic Language Resources Kits (BLARK) for the language.

Keywords: Natural Language Processing (NLP); Igbo Language; Text Analyzer.

1. Introduction

As a fundamental element of human communication, language embodies a wide scope of spoken dialects in the world. It comprises a system of symbols used by humans to communicate or express ideas and thoughts with one another (Abu-Rabia, 2012). Based on regions and tribes, there are numerous languages spoken globally, and as such, the Igbo language is among them. Igbo Language is one of the three major languages in Nigeria, alongside Hausa and Yoruba languages. It belongs to the Benue-Congo group of the Niger-Congo family with over 25 million speakers in Nigeria, and it is significantly acknowledged because of its rich linguistic structure, historical significance, and cultural heritage (Eberhard *et al.*, 2019).

Like any other language in existence, the Igbo language contains tones and vowel harmony characteristics. There are about thirty dialects in the Igbo language, with each having a peculiar contrastive pitch. Though the dialectal variations of the dialects are usually lexical, phonological, and syntactic, Emenanjo (1978) stated that the dialectal variety was not based on any particular place but rather, as a result of choosing the finest styles or ideas of the Igbo language from a range of available options, thus, the Owerri and Umuahia dialects spoken by indigenes of the two eastern states of Imo and Abia states in Nigeria became the Igbo language standard dialect (UCLA, 2014).

* Corresponding author: Prince C. Azubuike

In order to maximize the impact and reach of the Igbo language in the global world, a need to improve the language speed analysis and translation became of utmost importance. Sequel to this, artificial intelligence (AI) has aided in the construction of resources and tools that improved the investigation and understanding of natural languages in our present age of digital information and technological advancements. It is the science and engineering of making intelligent machines. Over the years, it has contributed to numerous significant progress in the various aspects of life in our society. Its branches include machine learning, expert systems, computer vision, cognitive computing, natural language processing, neural networks, robotics, and deep learning. Presently, the fields of NLP and computational linguistics is increasingly explored as the best methods for representing and analyzing human languages and developing resources and tools for low-resource languages (Khurana *et al*., 2022).

NLP is a disc approach that focuses on enabling computers to understand, interpret, and generate human language in a way that is similar to how humans communicate with each other (Khurana et al., 2022). It has spread its applications in various fields such as automatic summarization, sentiment analysis and opinion mining, coreference resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition, optical character recognition, part of speech tagging, low-resource NLP, etc. (Khurana et al., 2022).

In order to improve the function of understanding and generating language text, NLP is classified into two parts as follows: Natural language understanding (NLU) and Natural language generation (NLG) (Khurana et al., 2022). To understand human language, NLU algorithms are applied at the different linguistic levels, which include phonology (arrangement of sounds), morphology (word formation), syntax (language sentence structure), semantic syntax, and pragmatics (understanding). On the other hand, NLG is geared towards producing phrases, sentences, and paragraphs that are meaningful from an internal perspective.

In the area of Igbo NLP, has been done in the development of Igbo language resources and tools. Significant progress has been made recently; such efforts are works done in Igbo-part-of-speech tagging (Onyenwe et al., 2019), Igbo diacritic restoration (Ezeani et al., 2016), Igbo embedding-based analogy and similarity (Ezeani et al., 2018), Igbo machine translation (Ezeani et al., 2020), Igbo text summarization (Mbonu et al., 2022), and Igbo name entity recognition (Chukwuneke et al., 2022). However, there is still a lot to be done in building a robust digital resource for the Igbo language, which is why this research is important.

This paper presents an automated web-based Igbo text analyzer. This work is part of ongoing research to create a robust Igbo text analyzing system that will perform Igbo text statistical and readability analysis, POS tagging, Igbo text morph analysis and sentence summarization.

1.1. Igbo language

The Igbo language official orthography is known as Ọnwụ orthography. It was adopted and standardized by the Ọnwụ Committee in 1961. Uchechukwu (2008) showed that in the 1500s, before the Ọnwụ Committee, the Igbo tribe had a writing system called Nsibidi that was based on ideograms utilized by some secret cults, like the Ekpe and Okonko, for secret communications. The Ọnwụ standard orthography of Igbo is made up of 36 graphemes, which are the twenty-eight (28) consonants: b, gb, ch, d, f, g, gh, gw, h, j, k, kw, kp, l, m, n, nw, ny, n, ñ, p, r, s, sh, t, v, w, y, z, and the eight (8) vowels that are divided into two harmony groups based on Advanced Tongue Root (ATR) as ị, ụ, a, ọ (-ATR) and i, u, e, o (+ATR). Amongst the consonants are nine (9) digraphs: ch, gb, gh, gw, kp, kw, nw, ny, sh (Ọnwụ Committee, 1961; Onyenwe 2017). The vowels of the two harmony groups are combined according to vowel harmony to form Igbo words (Ọnwụ Committee, 1961; Emenanjo, 1978). For example, -ATR will have ịmi 'nose', ụlọ 'house', anya 'eye', ọbara 'blood', and +ATR will have igba 'drum', uwe 'clothe', ego 'money', oyi 'cold'.

The Igbo language is written using two-tone marking methods: the level tone and the level/gliding tone mark. Only contrastive tones are marked in the first system (Nwachukwu, 1987), while in the second system, all low tones are marked, leaving all high tones (Emenanjo, 1978). Also, using the second marking system, at the lexical level, the word 'egbe' when written without tone marks could mean gun or kite. These equivalents can be properly distinguished when tone marks are included as follows: égbè for gun and ègbe for kite.

2. Related works

Yao-Ting et al. (2015) developed a tool for the automated analysis of simplified and traditional Chinese texts called the Chinese Readability Index Explorer (CRIE). It has four subsystems and 82 multilevel linguistic features. The main tasks of CRIE were segmentation, syntactic parsing, and feature extraction. The integration of linguistic features with machine

learning models enabled the system to level and diagnose information for texts in language arts, texts for learning Chinese as a foreign language, and texts with domain knowledge.

Mikhail (2015) worked on a morphological analyzer and generator for Russian and Ukrainian languages called Pymorphy2. The system uses large, efficiently encoded lexicons built from Open Corpora and Language Tool data. Pymorphy2 provides state-of-the-art morphological analysis quality. The analyzer was implemented in the Python programming language with optional C++ extensions. Distributing the package under a permissive open-source license encouraged its use in both academic and commercial settings.

Itisree *et al*. (2015) designed a fully-fledged morphological analyzer (MA) tool for Oriya, which is an agglutinative language. The system was developed using the paradigm approach. The paradigms were created for inflected forms using an XML-based morphological dictionary from the Lttoolbox package. Presently, the dictionary contains 10,480 words.

Alexei *et al*. (2016) developed a morphosynthetic analyzer for the Tibetan language. The system creates a consistent formal grammatical description of the Tibetan language, ranging from morphosyntax (syntactics of morphemes) to the syntax of composite sentences and supra-phrasal entities. The syntactic annotation was created on the basis of a morphologically tagged corpus of Tibetan texts.

Samarjeet *et al*. (2017) developed a successful morph analyzer for non-declinable adjectives in Nepali. They developed the technique using a finite-state grammar approach, which can operate with a minimal number of linguistic resources.

Ekaterina *et al*. (2018) worked on an open source cross-platform morphological analysis library for the Russian language. They designed the library to function in multi-threaded applications with minimal performance loss and to incorporate additional data integrity controls into industrial high-load systems of any type. The system is very useful for linguists and software developers working on morphological analysis or word generation.

Zhandos *et al*. (2020) analyzed a pipeline for automatic processing of texts written in Kazakh. The system offered pre-processing tools such as text normalization and language identification.

Teodora *et al*. (2020) developed a syntax analyzer for the Serbian language. The system was based on context-free grammar. Firstly, the system describes building a POS tagger for the Serbian language and then defines context-free grammar for the Serbian language. The syntax analyzer was implemented using NLTK tools.

Darkhan *et al*. (2021) designed and developed a linguistic resource and pre-processing tool for the Kazakh language. The system consisted of three automatic text pre-processing tools for the Kazakh language: word form generation, a morphological analyzer, and a morphological disambiguation tool. The media corpus of the Kazakh language, which comprised the texts with news content, served as the foundation for the system's construction. Other applications of the system include automatic text analysis systems and the development of linguistic resources like thesaurus and ontologies.

3. Research methodology

Object-Oriented Analysis and Design Methodology (OOADM) was adopted for easy implementation of the proposed system. OOADM encourages the reuse of designs and components and gives users a better understanding of the program. The separation of different system components made it possible for reusability and parallel development, which significantly reduced the amount of time used in developing the system. Figure 1 shows the Data Flow Diagram of the proposed system.

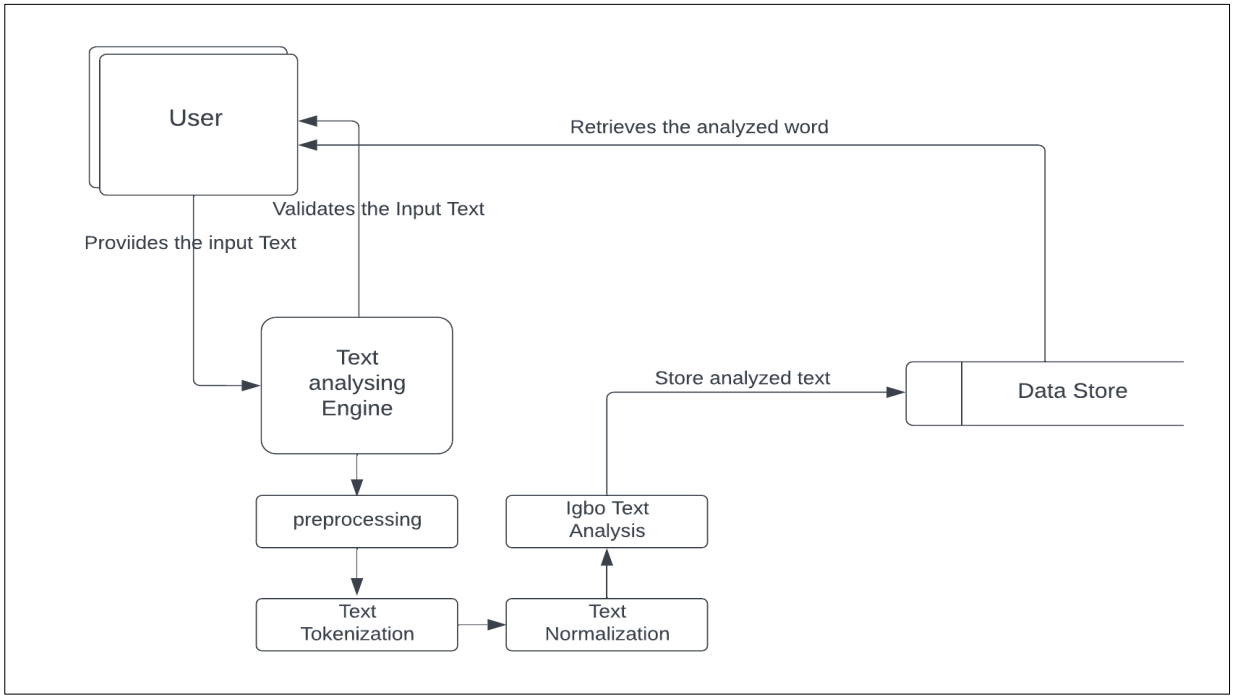

Figure 1 Data Flow Diagram of the Igbo Text Analyzer

4. Materials and method

4.1. IgbTC

The proposed Igbo text analyzer adopted the first Igbo-tagged corpus (IgbTC) developed by Onyenwe (2017). The new system adopted text analytics techniques as a feature to enhance the automatic analysis of Igbo text. The new system architecture employs Representational State Transfer (REST), also known as REStful web services, and Uniform Resource Protocol (URI) to respond to various document formats such as Extensible Markup Language (XML), Hypertext Markup Language (HTML), JavaScript Object Notation (JSON), and other light-weight data interchange-defined formats.

4.2. Natural Language Development Toolkit (NLTK)

NLTK and some other Python libraries were used to perform the analysis of the Igbo text that included word tokenization, sentence tokenization, and lemmatization.

4.3. UML Diagram of the Igbo Text Analyzer

Three Uml diagrams were used to analyze the system functionality. They are the Use case, Class and Activity diagrams.

4.4. The Use Case diagram

This was used in the design stage to show the actors in the system and the role they play, as depicted on Figure 2.

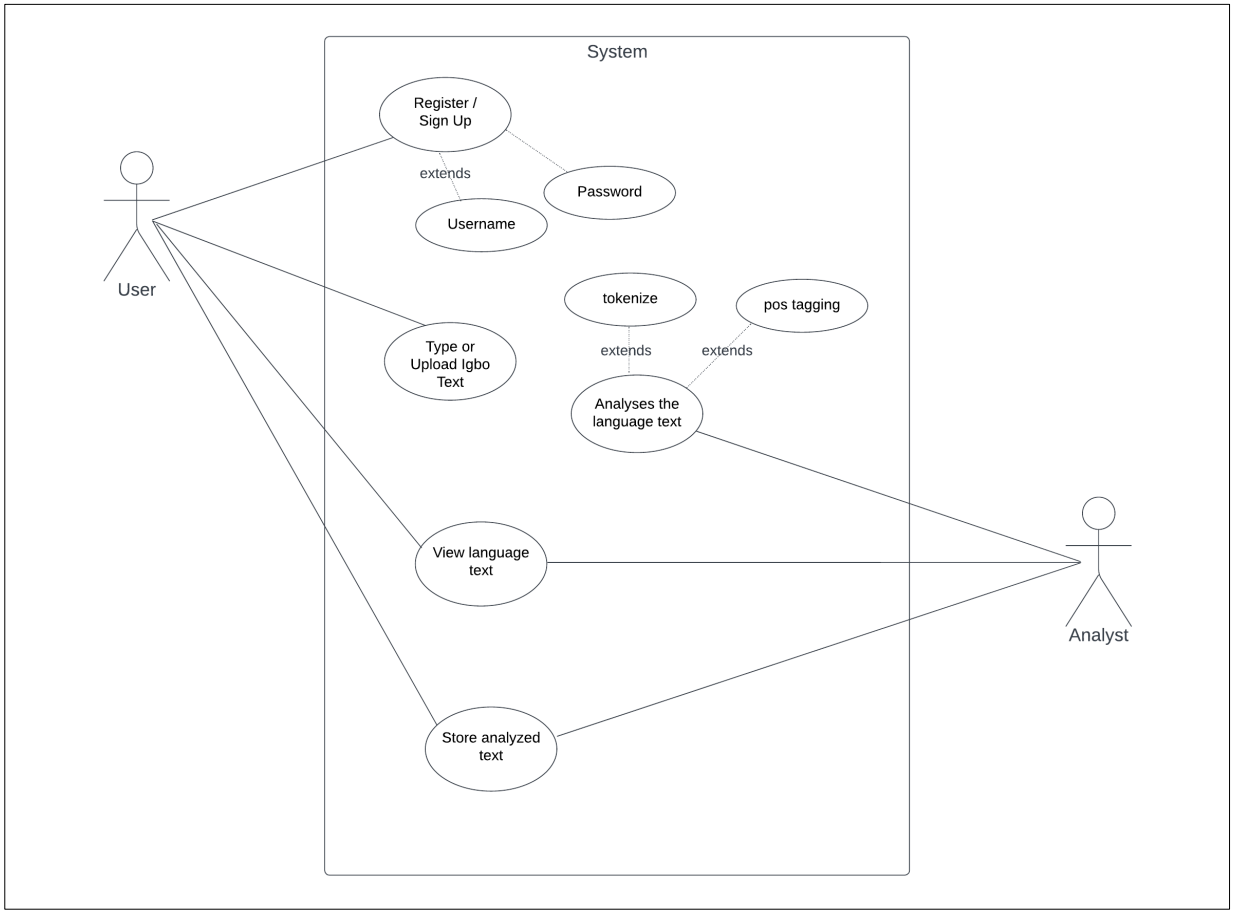

Figure 2 Use Case Diagram of the Igbo Text Analyzer

4.5. Class Diagram of the Igbo Text Analyzer

This shows the static behavior of the developed system. Figure 3 presented the static view of the application and described the attributes and operations of the object classes.

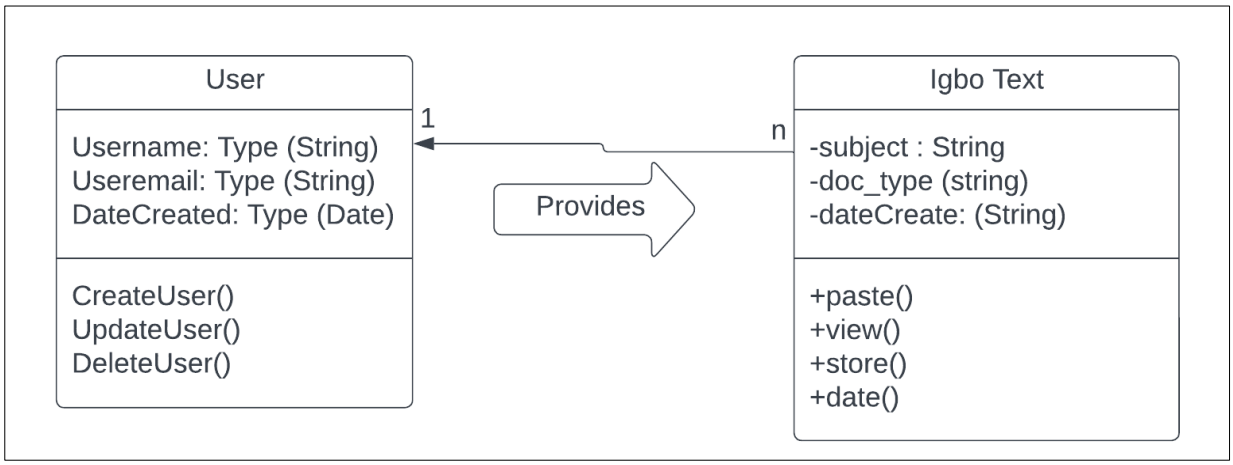

Figure 3 Class Diagram of the Proposed System

4.6. Activity Diagram of the Igbo Text Analyzer

The activity diagram represents the flow from one activity to the other in the system as shown on Figure 4.

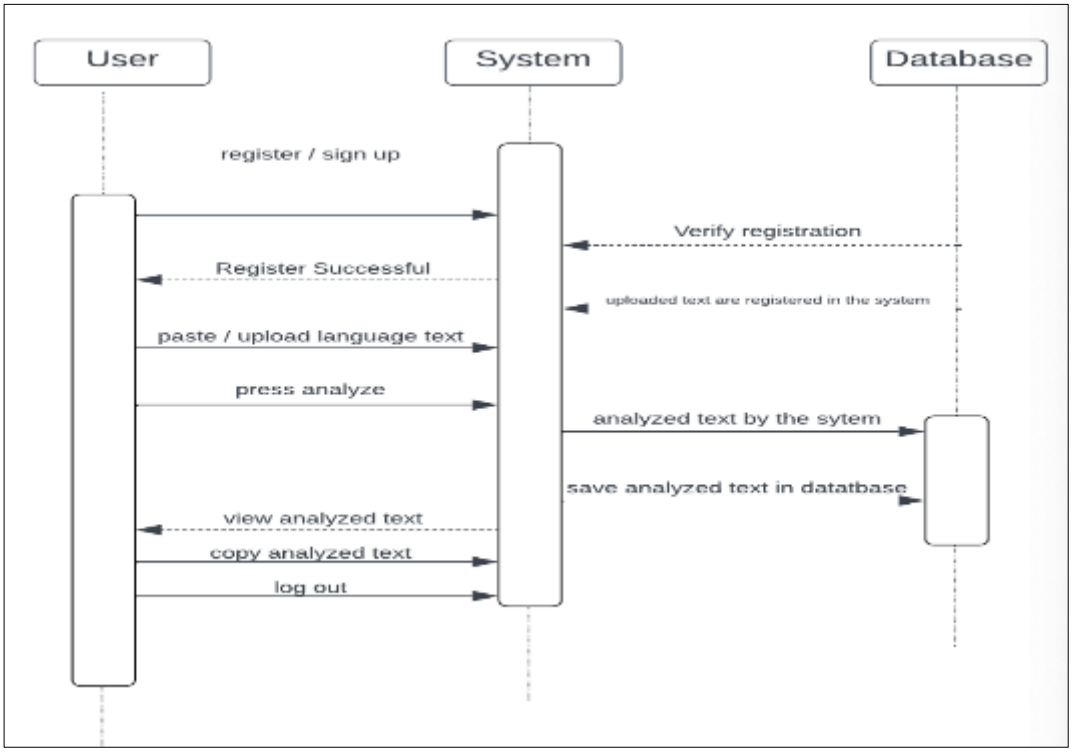

Figure 4 Activity Diagram of the Proposed System

4.7. Sytem Flow Chart

The system flowchart illustrates the flow of data within the system and the decision-making process involved in controlling events, as depicted on Figure 5.

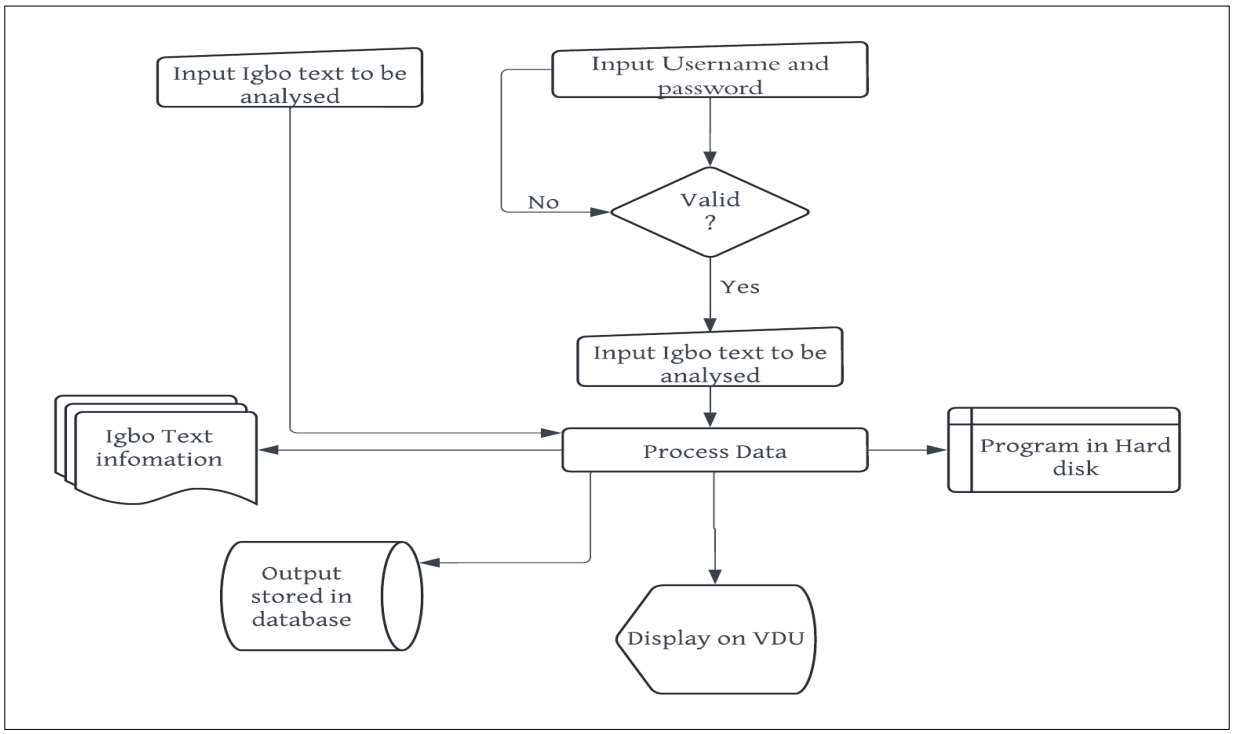

Figure 5 System Flow Chat

5. Results and discussion

The sampled results of the developed system are depicted on Figures 6-9, which show the login page, registration page, analysis input, and analysis output, respectively.

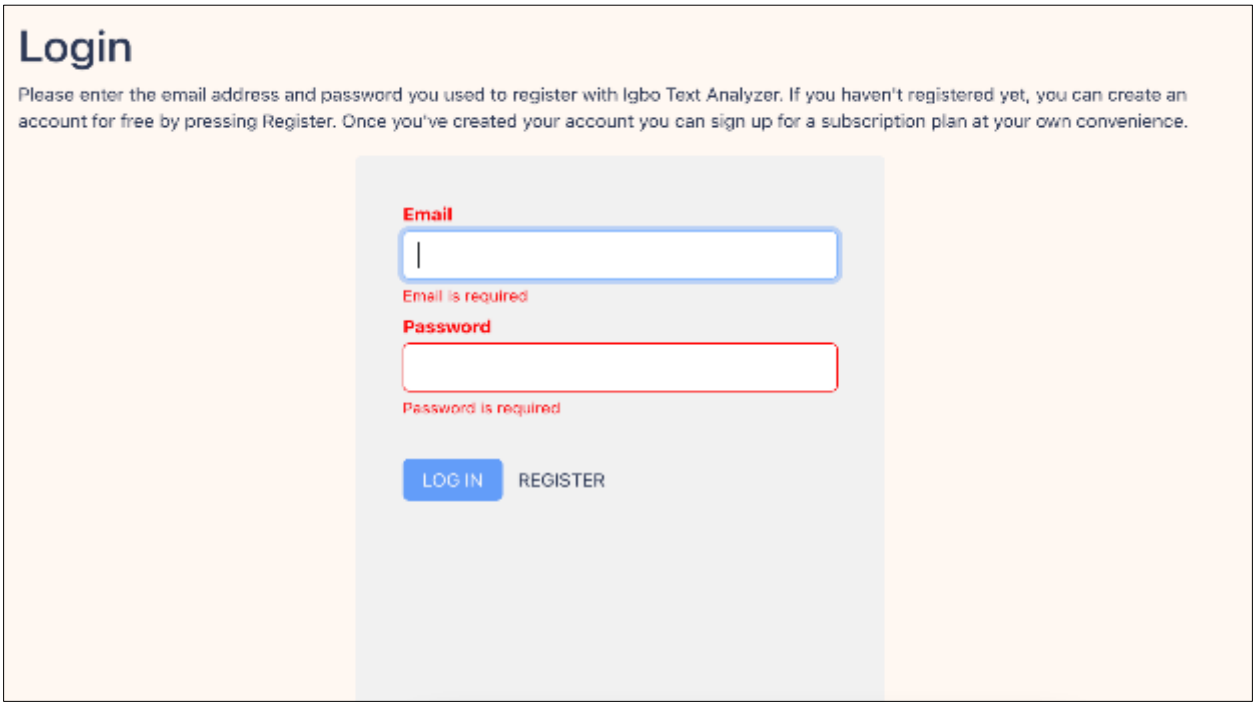

Figure 6 Login Page

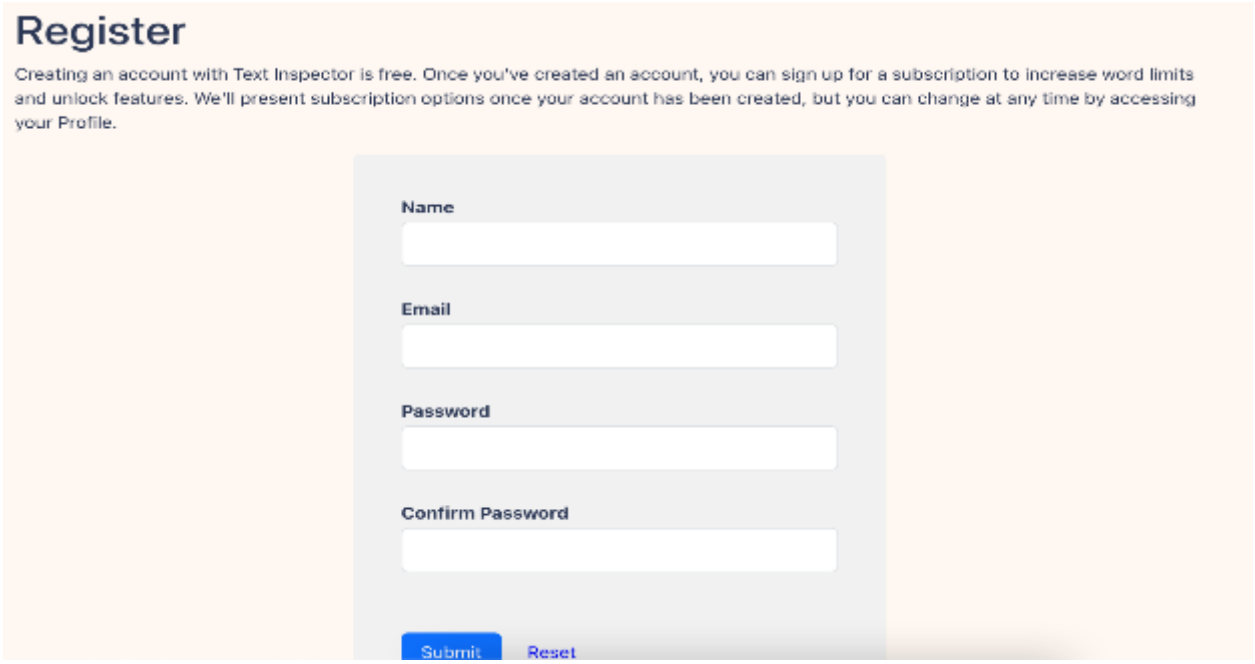

Figure 7 Registration Page

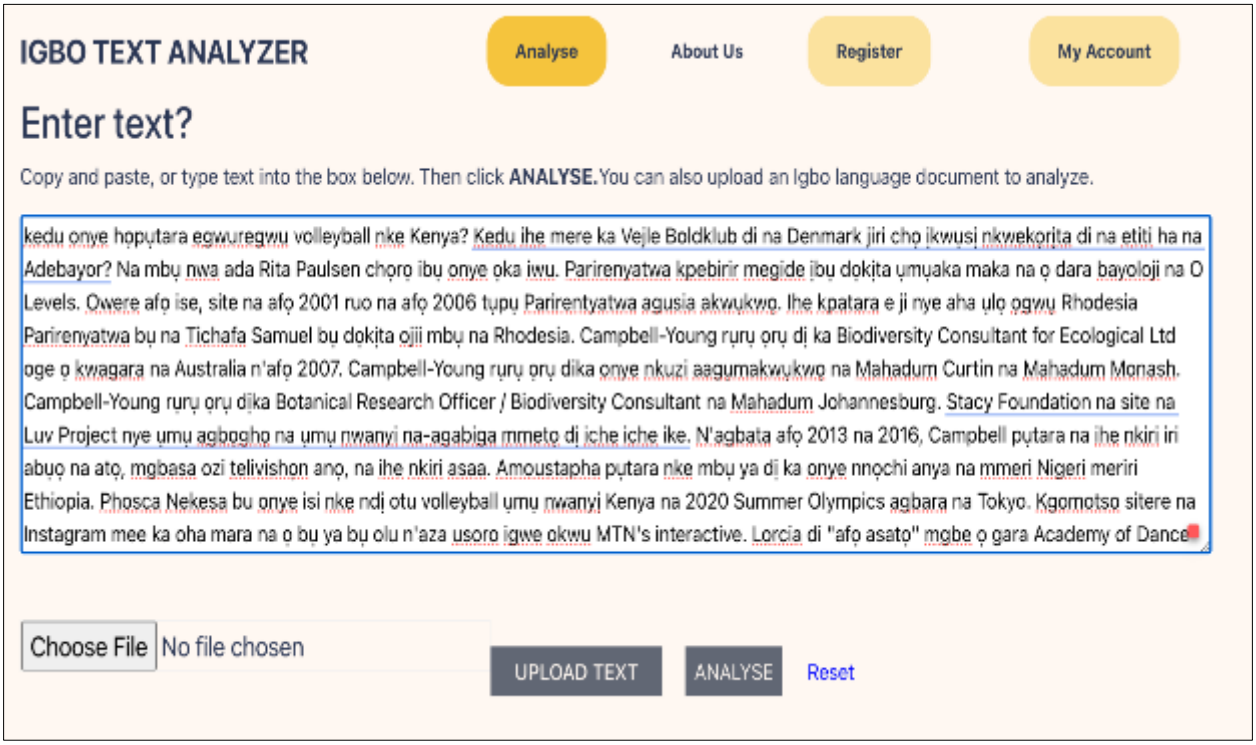

Figure 8 Analysis Input

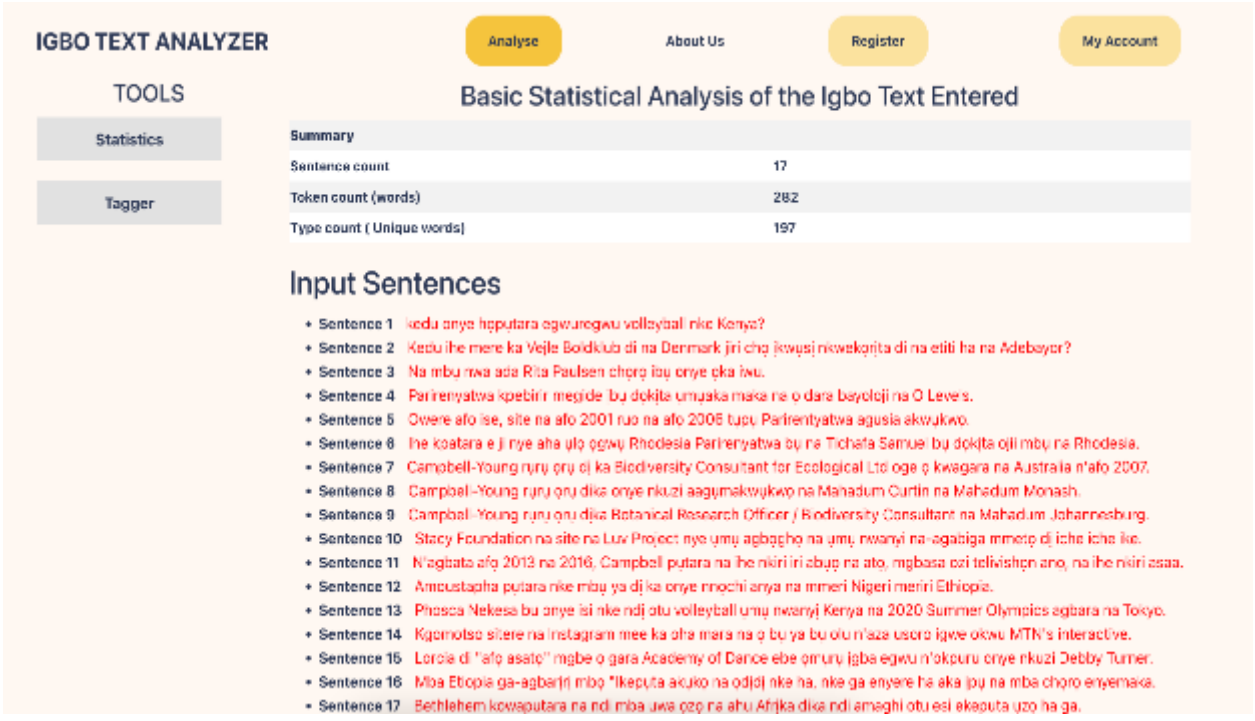

Figure 9 Analysis Output

6. Conclusion

The significance of this study cannot be overemphasized, particularly because it utilized several advanced technological tools to address the shortcomings in the automatic analysis of Igbo language text. Though the study revealed a

significant progress in the field of Igbo natural language processing, it still lacked in public visibility, stemming from the researchers' failure to integrate their findings with contemporary technology. The development of an automatic web-based Igbo text analyzer alleviated visibility challenges faced by the Igbo language NLP.

Compliance with ethical standards

Acknowledgments

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

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Abu-Rabia, Dana B. D. (2012). A Study into the Results of an Intervention Program of Linguistic Skills in English (L2) and its Effect on Hebrew (L1) among Poor Readers: An Examination of the Cognitive Retroactive Transfer (CRT) Hypothesis. Open Journal of Modern Linguistics, Vol. 4, No. 2, pp. 131-139.

[2] Alexei D., Anastasia D., Pavel G., Nikolay S. & Victor Z. (2016). Morpholsyntactic Analyzer for the Tibetan Language: Aspects of Structural Ambiguity. *International Conference on Text, Speech, and Dialogue,* Vol. 9924, *pp. 215-222.* DOI: 10.1007/978-3-319-45510-5_25.

[3] Darkhan A., Madina M., Gulmira M., Nurgali K., & Marzhan K. (2021). Development of the information system for the Kazakh language preprocessing. *Cogent Engineering*. Vol. 8. https://doi.org/10.1080/23311916.2021.1896418.

[4] Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). Ethnologue: Languages of the world. sil international.

[5] Emenanjo, E. N. (1978). Igbo Metalanguage (Okasusu Igbo). A glossary of English Igbo technical terms in language. Literature and methodology. Ibadan: University Press Ltd.

[6] Ekaterina V. P., Sergey A. P., & Alexander S. P. (2018). Development of the Cross-platform Library of Morphological Analysis of the Russian Language Text for Industrial Software. Proceedings of the 14th Central and Eastern European Software Engineering Conference, Russsia, Vol. 11, *pp. 1-8*. https://doi.org/10.1145/3290621.3290635

[7] Ezeani I., Mark H., & Onyenwe I. (2016). Automatic restoration of diacritics for the Igbo language. *In International Conference on Text, Speech, and Dialogue, pp. 198-205. Springer.*

[8] Ezeani I., Rayson P., Uchechukwu C., Mark H., & Onyenwe I. (2020). Igbo-English machine translation: An evaluation benchmark.

[9] Itisree J., Himani C., & Dipti M. (2015). Oriya Morphological Analyzer Using Lttobox. DOI: 10.13140/rg.2.1.2638.1520.

[10] Khurana D., Koli A., Khatter K., Singh S. (2022). Multimedia Tools and Applications. *Natural language processing: state of the art, current trends and challenges.* 82:3713-3744.https://doi.org/10.1007/s11042-022-13428-4

[11] Mbonu C., Chukwuneke C., Paul R., & Onyenwe I. (2022). IGBOSUM: Introducing the Igbo Text Summarization Dataset. *AfricaNLP Workshop at ICLR.*

[12] Mikhail K. (2015). Pymorphy2: Morphological Analyzer and Generator for Russian and Ukrainian Language. International Conference on Analysis of Images, Social Networks, and Texts, Vol. 542, *pp. 320-332.* DOI: 10.1007/978-3-319-26123-2_31.

[13] Onyenwe, I. E. (2017). Developing Methods and Resources for Automated Processing of African Language Igbo. *University of Sheffield Conference Proceedings.*

[14] Onyenwe I. E., Mark H. Uchechukwu C., & Ignatius E. (2019). Toward an effective Igbo part-of-speech tagger. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). Vol. 18, No. 4, *pp. 1-26.*

[15] Samarjeet B., Upali C., & Nermit L. (2017). *Design of a Morp Analyzer for Non-Declinable Adjectives for Nepali Language*. Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, pp. *126–130.* https://doi.org/10.1145/3036290.3036307.

[16] Teodora D. & Suzana S. (2020). A Tool for Sentence Syntax Structure Markup for the Serbian Language.

[17] Uchechukwu, C. (2008). African language data processing: The example of the Igbo language. In the 10th International Pragmatics Conference, data processing in African languages.

[18] UCLA (2014). Language materials project: Igbo. Accessed:2023-08-31

[19] Yao-Ting S., Tao-Hsing C., Wei-Chun L., Kuan-Sheng H., & Kuo-En C. (2015). CRIE: An automated analyzer for Chinese texts. *Behaviour Research Methods.* Vol. 48, pp. 1238-1251. https://link.springer.com/article/10.3758/s13428-015-0649-1

[20] Zhandos Y., Zhanibek K., & Aibek M. (2020). KazNLP: A Pipeline for Automated Processing of Texts Written in Kazakh Language. *International Conference on Speech and Computer (SPECOM 2020).* Vol. 12335. DOI: 10.1007/978-3-030-60276-5_63.