University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 15 No. 1, September 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

Igbo Text Named Identity Recognition (NER) System using Natural Language Processing Algorithms: A Review

[1]✉*Jacinta Chioma Odirichukwu [2]Precious Kelechukwu Chika-Ugada,, [3]Reginald Nnadozie Nnamdi, [4]Simon Peter Chimaobi Odirichukwu, [5]Chinwe Ndigwe, [6]Oluwatobi Wisdom Atolagbe, [7]Chigozie Dimoji, [8]Obilor Athanasius Njoku, [9]John Chinenye Nwoke, , [10]Godwin Oko Ekuma, [11]Iyanu Tomiwa Durotola, [12]Chiedozie Raphael Dunu, [13]Joshua Nzubechukwu Dinneya, [14]Felix Nmesoma Diala, [15]Samuel Chizitaram Dialaeme-Diolulu, [16]Chukchukwuka Prince Liberty, [17]John Prince Uzodinma, [18]Ezekiel Gabriel Nwibo

[1,2,7.8,12,13,14,15,16,17]Department of Computer Science, Federal University of Technology, Owerri (FUTO)
[3]Department of Philosophy, Veritas University Abuja.
[4]Department of Health, Primary Health Development Agency, Owerri, Imo State, Nigeria
[5]Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University (COOU), Uli
[6]EOS Energy Storage, Edison, NJ, USA
[9]CISCO/ICT Unit, Federal Government College, Port Harcourt, Rivers STate, Nigeria
[10]Department of Computer Science , Missouri State University, Springfield, MO, USA
[11]Department of Computer Science, Maharishi International University, Fairfield, IA, USA
[18]Department of Computing, School of Arts and Creative Technologies, University of Greater Manchester Bolton, UK.

[1*]jacinta.odirichukwu@futo.edu.ng; chiomajaco6@gmail.com

*Correspondence Author

Abstract
This is a review paper, which is concerned with the recent nature of Named Entity Recognition (NER) for the Igbo language. It is a low-resource language spoken in the Southeastern part of Nigeria. Irrespective of the numerous advancements in NER for high-resource languages, Igbo NER so far remains underrepresented. This is for its unique linguistic challenges, which includes morphological richness and dialect variations. In recent times, frank efforts have been put forward by MasakhaNER and WAZOBIA NER projects to develop NER datasets and models for the Igbo language. The existing datasets are limited in size and domain coverage. For this reason there are needs for high-quality, large-scale, manually annotated NER datasets for real-world deployment. This paper reviews the existing literature works on Igbo NER, highlighting the challenges, creating opportunities and looking into the potential applications of NER in developing Igbo digital assistants, intelligent search, and machine translation. This work aims to contribute to the growth and development of low-resource African NLP with the provision of future research in indigenous language NER.

Keywords: Named Entity Recognition, Igbo Language, Low-Resource Languages, Natural Language Processing, African NLP

1. Introduction

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP) which involves automatically identifying and classifying named entities like persons, locations, organizations and so on within unstructured text. It is the foundation for different NLP applications such as information extraction, machine translation, question answering, and intelligent search. Although

Jacinta Chioma Odirichukwu, Precious Kelechukwu Chika-Ugada,, Reginald Nnadozie Nnamdi, Simon Peter Chimaobi Odirichukwu, Chinwe Ndigwe, Oluwatobi Wisdom Atolagbe, Chigozie Dimoji, Obilor Athanasius Njoku, John Chinenye Nwoke, Godwin Oko Ekuma, Iyanu Tomiwa Durotola, Chiedozie Raphael Dunu, Joshua Nzubechukwu Dinneya, Felix Nmesoma Diala, Samuel Chizitaram Dialaeme-Diolulu, Chukchukwuka Prince Liberty, John Prince Uzodinma, Ezekiel Gabriel Nwibo (2025). Igbo Text Named Identity Recognition (NER) System using Natural Language Processing Algorithms: A Review. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 15 No. 1, pp. 13 - 21
©U IJSLICTR Vol. 15, No. 1, September 2025

NER systems have been really developed for high-resource languages like English, German and Chinese, low-resource languages like Igbo remains underrepresented. The Igbo language, widely spoken in southeastern Nigeria has unique linguistic challenges for NER research because of its morphological richness, different dialects and a lack of standardized annotated corpora. Some efforts have been made in recent years to address these challenges. MasakhaNER project was a pioneering multilingual initiative at NER dataset development for 10 indigenous African languages, the Igbo language inclusive. It used manually sourced annotation to create its annotated corpora and multilingual transformer models (mBERT and XLM-R) on Africa NER tasks [1].

After MasakhaNER, the WAZOBIA NER project was developed using Nigeria's three major languages - Igbo, Hausa, Yoruba [2]. As the Igbo NER dataset in this project was very small. [3] introduced a projection approach to expand Igbo NER datasets by lining named entity annotations from English to Igbo using parallel corpora. This increased dataset size and improved training and generalization of NER models across domains. On another side, the IgboBERT models were introduced as domain transformer models that were pre-trained on Igbo corpora and fine-tuned for NER tasks. This achieved promising results as compared to multilingual baselines [4]. Another important contribution to the Igbo NLP ecosystem is the IgboAPI dataset that offers a structured and enriched linguistic resource designed to support the development of various Igbo language technologies, including NER [5]. The dataset has orthographic normalization, part-of-speech tagging, and lexical entries across dialects that are important for developing robust NER systems that generalize well across the Igbo region.

Despite all these advancements, significant gaps remain. The existing Igbo NER datasets are limited in size, or rely heavily on news and Wikipedia data, or are not rich in dialect variations. Also, minimum efforts have focused on building high quality, large scaled, manually annotated NER datasets specifically for real-world deployment.

Named Entity Recognition (NER) has really advanced use in high-resource languages, but the same cannot be said about low-resourced languages like Igbo. This is because of the lack of sufficient annotated corpora to train and evaluate these systems. There are very few existing Igbo NLP dataset, yet they are small in size, limited in domain coverage, and lack proper entity annotations. NER models trained on such data with these limitations would not generalize effectively across domains such as news, education, healthcare and social media.

The current NLP pipelines for the Igbo language depend heavily on rule-based approaches or zero-shot transfer learning. This often leads to poor performance when handling context-sensitive or culturally unique named entities. So, there is a clear need for an accurate and scalable NER system suitable for the structure and syntax of the Igbo language.

2. Related Works

2.1 Conceptual Framework

2.1.1 Definition of Named Entity Recognition (NER) and its Importance

Named Entity Recognition (NER) is a technique used to identify and label the names of people, places, organizations, and other well defined items in text. NER enables systems to mark important information in such a way that it is easier to work with. This is important because tasks like summarization, question answering, and search mostly rely on being able to identify names within a document and understand how those names are used. This step comes before more complex interpretation can happen.

Harrando [6] showed that recognizing these entities makes it easier to summarize content from complex sources like videos or audio transcripts. In another study, Muppavarapu *et. al.* [7] used entity recognition to connect related ideas in Web of Things data by comparing concepts that were semantically close. Chatterjee [8] applied similar methods in building systems that understand search queries by focusing on the entities users include in their questions.

2.1.2 Low-Resourced Languages and the Igbo Language

Languages are considered low-resource when they do not have sufficient labelled datasets or a clear linguistic documentation. In Nigeria which has over 500 languages, the three major languages, Hausa, Igbo, and Yoruba still fall into this category. Although these three are

used by millions daily, they remain underrepresented in the field of natural language processing. Inuwa-Dutse [9] observed that only a few studies in the Nigerian context have focused on developing new datasets. Most researchers rely on existing data which implies that the Igbo language has not seen much progress in terms of corpora and/or NLP tools.

Usip *et. al.* [10] examined the limitations when building an emergency alert system for languages in the Southern region of Nigeria. Their findings showed that Igbo and Ibibio lacked enough digital resources to support even simple rule-based tools. Ekle & Das [11] trained translation models on English and Igbo texts drawn from Bible verses, news reports, and Wikipedia entries. They used transfer learning to train and raised the BLEU score to just under 70. This result goes on to imply that cleaner and a well prepared dataset could push the performance even higher. But while this is promising, most of the current efforts still focus on translation, leaving NER tasks behind.

2.1.3 NER Labelling Schemes

When developing a named entity recognition system, selecting a tagging scheme is an important first step. The two widely used formats are BIO (Begin, Inside, Outside) and BILOU (Begin, Inside, Last, Outside, Unit). These tags provide structure on how the model sees each word in relation to named entities. For instance, BIO tags help the model tell when an entity begins and continues, while BILOU adds more detail by marking the last token and single-word entities separately. The choice between the two can affect how much the system learns entity boundaries. One study on weak supervision observed that using a structured tagging scheme together with focused handling of uncertain labels made it easier for the system to learn from noisy dataset [12].

In another case, researchers working with texts from the health domain applied the BIO format while using recurrent models and found that it helped the system mark and classify terms more accurately [13]. In another study, BILOU tagging was applied during the extraction of entities and their relations. This made the model define clear boundaries around each entity and supported more accurate identification of how the entities were linked within the text [14].

2.1.4 Challenges of NER in Low-Resource and Morphologically Rich languages

The major challenge in NER for under-resourced languages like Igbo is the absence of detailed language descriptions or parsing tools.

Chaudhary [15] made the AutoLEX system to automatically extract linguistics properties like morphology, syntax and lexical semantics from raw text. She discovered that without these descriptions, it was nearly impossible to create reliable tokenizers and POS taggers. In relation to Igbo, efforts to build NER systems face the same problem of limited resources to enable building of foundational language analysis tools.

Tedeschi [16] developed a framework that revealed that tools designed for high-resource languages often fail in their low-resource counterparts because specific phenomena like inflection, compounding, complex agreement patterns were ignored. Similarly, Jembere [17] while working on the Amharic language revealed that universal dependency parsers struggle when faced with morphologically rich languages and inconsistent annotations. This becomes a real barrier for building effective Igbo NER systems.

2.1.5 Role of Pretrained Language Models in NER

Pretrained language models have made it easier to build NER systems for languages that lack large datasets. BERT and XLM-R are two examples that have been widely used. These models are first trained on general text in many languages and then fine-tuned on a specific task, like NER, using smaller, labelled datasets. Jiang *et. al.* [18] introduced XLM-K, which adds multilingual knowledge from sources like Wikipedia and Wikidata into the pretraining stage. This helped the model better understand links between languages and improved how well it handled named entities in different languages.

Jean [19] showed that multilingual pretrained models can support low-resource languages without the need to start building from scratch. When fine-tuned on the local data, they adapt to the specific language patterns thus improving its accuracy. Kalyan *et. al.* [20] also looked at several pretrained models and found that even though BERT and its variants are general-purpose, they can work well for NER if the fine-tuning is done carefully.

2.2 Theoretical Framework

2.2.1 Sequence Labelling and Token Classification Theory

Named Entity Recognition is based on the idea of sequence labelling. Here, each word in a sentence is tagged separately, depending on whether it is part of a named entity or not. Some words might begin an entity, others may continue it, and some are not part of any entity at all. This is not like sentiment analysis, where the model simply gives one label to the whole sentence. In NER, each word is looked at on its own, and the model tries to figure out what role it plays in the sentence.

Rigouts *et. al.* [21] noted that sequence labelling performs well for tasks such as term extraction because it enables the model to learn from how words appear together in a sentence. Rather than focusing on words in isolation, the model is guided by patterns found in their arrangement.

Jafari [22] looked at the difference between token classification and sequence classification. He found that token-based methods work better when the task involves tagging each word separately, like in NER. This makes sense, since one sentence can contain several names, each with a different label.

Arora *et. al.* [23] applied token-level sequence labelling in their work on spoken language understanding. Although their study focused on speech rather than text, the approach proved effective. It allowed the model to handle unclear inputs more reliably by making decisions at the level of individual words.

2.2.2 Transformer-based Architectures

BERT and XLM-R are models that are now used for tasks like Named Entity Recognition. They use a concept known as self-attention. What this means is that instead of reading the sentence word by word, the model looks at everything together. It becomes easier to see how the words connect.

Chen *et. al.* [24] looked at how the attention part of the model works. They found that the regular setup tends to take up a lot of GPU memory, especially with longer inputs. So, they changed the structure a bit and came up with a new one called ET. It was designed to reduce memory load while still keeping the model accurate during training.

Zuo & Mak [25] focused on sign language, but the idea in their study can apply here too. Their model gave more weight to nearby elements, instead of treating every input the same. That is useful for NER, where the words close to a name can often help decide what the name refers to.

Rahali & Akhloufi [26] reviewed how transformer models are used in language tasks. They pointed out that models like BERT already know a lot about language because of how they were trained. When fine-tuned for NER, there is no need to change much. A small classification layer is enough to label the tokens.

2.2.3 Transfer Learning and Domain Adaptation

Transfer learning is basically taking a model that has been trained on one task and applying it to a different but related task. The main idea is to make use of the patterns or representations the model has already picked up when starting from scratch is not practical. This is common in NLP where large labelled datasets are sometimes hard to find. In situations like Igbo NER, because annotated data is limited, transfer learning is very useful to reduce training time and improve performance.

Luo *et. al.* [27] applied cross-language transfer learning to speech recognition. They found that models trained on high-resource languages could still perform well on lower-resource ones when domain adaptation was used. Their results showed improved accuracy, even with limited data from the target language. That kind of method could also help Igbo NER, which faces a similar data scarcity issue.

Sometimes, the data used to train a model is not the same as the one it saw before. That is why Lu *et. al.* [28], even though worked on energy prediction, showed that it helps to mix transfer learning with domain adaptation. For a task like Igbo NER, where the text might be taken from news or blogs and not Wikipedia, this kind of approach can still work. Lin *et. al.* [29] also studied how to deal with differences between training and testing data. They worked on a non-text task, but the idea is similar. They used a network that could transfer what the model learned and adjust it to fit the new data. Even if the domains are not the same, it gave better results than just copying the model as it was.

2.2.4 Annotation Theory and the Semiotics of Named Entities

When working on NER, annotation is not just one technical step that could be ignored. It is about understanding what the words are doing in a sentence and what they actually mean. In Igbo where names often hold cultural, emotional, or historical meaning, this is important.

In Antonini & Brooker [30], names are not just tags for people or places but carry a memory, a meaning, and a real sense of identity. That is true in the Igbo language where names often tell a story or reflect something deeper about where someone is from or what they believe. Most generic models are not built to catch such. Bateman [31] added that documents are not just plain text holders. Rather, they carry meaning in different ways, structure, visuals, and language. In NER, named entities should not be treated as random labels. They are part of a wider system that builds meaning across different parts of the text.

Antia & Mafofo [32] supported this idea by saying annotation is something people naturally do. Whether it is through speech, writing, or body movement, people find ways to mark meaning. That implies something important. When working with NER in a language like Igbo, tagging cannot be done like a machine task. It has to reflect how people in that language and community actually use and understand names.

2.3 Empirical Review

2.3.1 Review of MasakhaNER Project

MasakhaNER is a multilingual dataset developed for named entity recognition tasks. It has ten African languages, including Igbo, Hausa and Yoruba. Adelani *et. al.* [1] addressed the lack of annotated data for these languages by producing labelled corpora, then adopted a consistent scheme across the languages. Models were trained with multilingual and unique language transformer based architecture. Results showed that the trained models, AfriBERTA and fine-tuned mBERT performed considerably well though performance was varied by language probably because of the differences in available text data and annotation quality.

The study faced some limitations. One, the entity category was restricted to four types (person, location, organization and date). This means that the range of named entities that could be captured was limited. For languages with rich cultural or context-specific references, it is not optimal. Another challenge was in domain generalization. Though the models performed well on data that were similar to training data, they still showed a noticeable drop in performance when applied to texts from other sources. Annotation faced difficulties in this project due to the spelling norms being inconsistent and then the absence of established linguistics tools. In spite of these, MasakhaNER remains a great step to improving NER systems for African languages and supports the larger goal of developing inclusive NLP technologies.

2.3.2 Development of WAZOBIANER system

The WAZOBIANER system is a worthy contribution to note in NLP research for under-resourced Nigerian languages. Emedem *et. al.* [2] built annotated datasets for Hausa, Igbo and Yoruba together thereby addressing data scarcity. They evaluated CRF with modern deep learning architectures (BiLSTM, and a BERT model improved with an RNN layer) to recognize person, organization and location entities. The inclusion of OCR processing allowed the system to work with both typed text and text from images. The F1-score reached 0.9564, suggesting that NER systems for low-resourced African languages can match strong benchmarks after careful model selection and preprocessing are obtained.

However, the NER system was limited to three entity types. So it could not be useful in domains that require dates, or numerical values. The study was focused on formal text sources leaving open questions about its performance on informal, and inputs with rich dialects. Also note that though OCR was infused in this system, the impact of OCR errors on the overall system accuracy was not discussed. Again, no experiment was conducted to assess possible cross-lingual transfer amongst Hausa, Igbo and Yoruba. Notwithstanding these gaps, the WAZOBIANER project has thoughtfully designed pipelines that can advance NER for languages that have long been overlooked.

2.3.3 IgboNER 2.0 Dataset

Chukwuneke *et. al.* [3] proposed IgboNER to expand named entity recognition resources for Igbo language using cross-lingual projection techniques. A parallel English-Igbo corpus was used to make a mapping dictionary linking

English entities to their Igbo equivalents. This dictionary was verified manually and then used to annotate an Igbo monolingual corpus. This brilliant idea resulted in a larger and robust NER dataset that improved downstream model performance, saving time and effort as compared to manual annotation. It also allowed for reuse in other NLP tasks.

The projection approach encountered some challenges such as missing segments in translations, inconsistencies in spelling and a lack of standardization in Igbo terms. Some English entities had no Igbo equivalent and multiple variants existed for the same word. This was worked on by updating the mapping dictionary to accommodate some of these word variations. IgboNER 2.0 underperformed slightly compared to larger multilingual models, but still showed that performance improves with increased data.

2.3.4 IgboBERT and transformer models for Igbo NER

IgboBERT is the first ever transformer model trained solely on Igbo text. Chukwuneke *et. al.* [4] collected and annotated a robust Igbo NER dataset before training a fresh language model from scratch. It was then fine-tuned using mBERT, XLM-R and DistilBERT. The experiment tested two learning rates, 1e-4 and 2e-5, and revealed that the multilingual models outperformed IgboBERT. For example, mBERT reached an F1 score of 89.02 and accuracy of 98.05 at 2e-5, while IgboBERT achieved a respectable 77.94 F1 score at 1e-4. The research also flagged possible overfitting with IgboBERT as training and validation losses failed to converge, an observation they could not fully address in that work.

Still, the experiment has lasting value. It shows that even with limited Igbo data, a dedicated transformer model can perform reasonably well in NER. The authors suggest that increasing pre-training data or using gazetteers could raise performance further. They have also released their code and models publicly, encouraging follow-up research.

2.3.5 The IgboAPI Dataset for Multidialect Processing

The IgboAPI dataset was developed in the absence of dialectally diverse resources for Igbo. The Igbo language which already is considered an endangered language. Emezue *et. al.* [5] made a practical contribution to Igbo language processing by compiling a multi-dialect Igbo-English dictionary. Two main experiments were used to evaluate the dataset, namely machine translation and semantic lexicon development. In machine translation, the authors fine-tuned an existing M2M-IBO-EN translation model using the IgboAPI data, resulting to an increase in BLEU scores from 16.87 to 71.95 on the standard test set and from 16.77 to 67.91 on unseen dialect test sets. The second experiment, semantic lexicon development was used to prove the utility of the dataset to enhance Igbo semantic tagging by mapping its definitions from English using PyMUSAS.

The study however, points out the difficulty of generalizing to dialects not represented in the training data, as there was reduced performance for variants like Egbema and Ubani. Ablation studies further confirmed that datasets containing dialect-specific features produced better outcomes. While performance varied across dialects, the authors note that these fluctuations were tied to the uneven distribution of training samples.

2.4 Summary of Literature Review

Despite progress in NER for Igbo and other low-resource languages, a major concern is the lack of large and varied annotated datasets. Most of the Igbo NER datasets reviewed are small, limited in the entity kinds they include, and drawn from narrow sources. Domains like social media, spoken conversations and so on are barely represented. Most models fell short in the area of dialects. IgboNER 2.0 and IgboAPI tried to solve this and expand datasets using English to Igbo translations and dictionaries, but errors in translations and word consistency were introduced. BERT, XLM-R and mBERT transformer models have performed well in Igbo NER tasks even with limited data. Although fine-tuning these models on the specific language datasets gives good results, the models still struggle when moved outside the training data.

WAZOBIA-NER still took a step further by adding Optical Character Recognition (OCR) to extend to text from scanned images for NER tasks, but the effect of OCR errors was not deeply tested. One future work from these studies is the need for better datasets that reflect more domains and dialects. There is also room to explore models trained specific to Igbo from scratch or improved existing ones after careful adaptation. Such tools will help improve the

accuracy and flexibility of Igbo NER systems and also for other languages in similar conditions.

3. Methodology

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) was used in the review procedure. The PRISMA 2020 guidelines were adopted for the purpose of transparency and reproducibility of the result. Four major stages of the methodological framework such as the identification stage, the screening stage, eligibility stage and the inclusion stage were all involved.

3.1 Search Strategy

Google Scholar was extensively used in carrying out thorough search, others were IEEE Xplore, ACL Anthology, SpringerLink, and arXiv preprints which spanned from 2020 to 2025. Some of the terms searched for include "Igbo NLP", "MasakhaNER", WAZOBIANER", "IgboBERT", "IgboAPI" and "Igbo Named Entity Recognition. To broaden the retrieval scope, Boolean operators (AND/OR) were comprehensively applied.

3.2 Inclusion and Exclusion Criteria

Inclusion criteria: Scholarly works published in English language ranging from 2020 to 2025 were focused on Igbo NER, dataset development, with related model training. More consideration was also put on peer-reviewed and preprint articles.

Exclusion criteria: On the exclusion list were Non-Igbo studies, duplicates, those work which are not related to NER (translation-only studies). More to this are those articles which lack empirical or insufficient methodological details.

3.3 Study Selection

Through the selection process, the initial search recorded 146 results. 48 duplicates were removed, afterwards 98 records were screened. Through the abstract and title screening, 42 full-text articles were fully assessed to ascertain eligibility. Again, 10 full-text were further excluded for not meeting inclusion criteria. 32 studies made it to the final synthesis. The flow chart in Figure 1 showed the selection process using the PRISMA process.

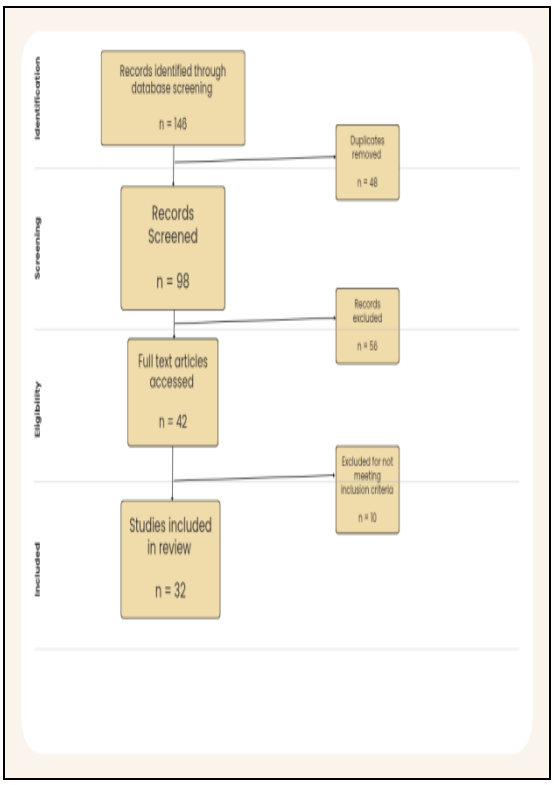

Figure 1: PRISMA Flow Diagram

3.4 Data Synthesis

The work came out through multiple processes, also multiple outcomes were found, a narrative synthesis was used rather than a meta-analysis or analytical procedure. The process gave room for wider discussion of dataset creation, the model performance, methodological gaps, and the opportunities which lie in future for the Igbo NER.

4. Results and Discussion

4.1 Overview of Included Studies

The PRISMA flow diagram (Figure 1) shows the step by step process of identification, screening, eligibility assessment, and inclusion. Out of 146 records initially identified, only 32 studies met the eligibility criteria and were included. These studies reflect ongoing but fragmented efforts toward developing Igbo Named Entity Recognition (NER).

4.2 Dataset Limitations

It was observed that most datasets found are small, domain-specific, which are found limited to a narrow range of entity types. The WAZOBIA-NER and MasakhaNER were the NLP that provided vital foundational resources though restricted entity categories. The IgboNER 2.0 and IgboAPI helped in expanding the coverage, but there are issues with

translation consistencies, orthographic variation, and uneven dialect representation; these impose challenges.

4.3 Model Performance

The Transformer-based multilingual models (mBERT, XLM-R, AfriBERTa) consistently outperformed Igbo-specific models like IgboBERT. The implication is that data scarcity limits the effectiveness of models trained solely on Igbo corpora.

4.4 Methodological Gaps

Some studies have shown that domain generalization has grown above some formal sources like Wikipedia and news. It is observed that informal texts, conversational data, and social media have been underrepresented. OCR error propagation was not addressed with the integration of WAZOBIA-NER's but the scope was broadened.

4.5 Opportunities for Advancement

The work has shown that opportunities abound in many ways on the Work by looking into the development of large-scale annotated corpora, hybrid approaches combining transfer learning with Igbospecific tools; expansion of entity categories; and creation of application-driven NER systems for Igbo language technologies.

4.6 Implications for African NLP

In today's world Igbo NER are faced with some challenges, such as representative of broader African low-resource NLP issues. Multilingual pretrained models provide strong baselines, for there to be sustainable progress, there should be well coordinated community-driven annotation, also dialect-sensitive resources and well opened non-restricted access dataset.

5. Conclusion

This review paper put forward some novel spots of Named Entity Recognition (NER) for the Igbo language, a low-resource language spoken in southeastern Nigeria. NER has proffered solutions to numerous challenges posed by unique linguistic characteristics, through recent efforts made in developing NER datasets and models for Igbo language. Irrespective of these strides, there are still needs for high-quality, large-scale, manually annotated NER datasets to support real-world deployment. The sole aim of the review is to contribute to the growing field of low-resource African NLP which helps in providing a foundation for future research in indigenous language NER. In addressing the challenges and opportunities in Igbo NER, one can unlock the potential of NER in developing Igbo digital assistants, intelligent search, and machine translation, ultimately promoting the preservation and development of the Igbo language.

6. References

[1] Adelani, D. I., Abbott, J., Neubig, G., Dsouza, D., Kreutzer, J., Lignos, C., ... &Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116-1131.

[2] Emedem, S. E., Onyenwe, I. E., &Onyedinma, E. G. (2025). Development of a WAZOBIA-Named Entity Recognition System. arXiv preprint arXiv:2505.07884.

[3] Chukwuneke, C. I., Rayson, P., Ezeani, I., El-Haj, M., Asogwa, D. C., Okpalla, C. L., &Mbonu, C. E. (2023). IGBONER 2.0: EXPANDING NAMED ENTITY RECOGNITION DATASETS VIA PROJECTION. In 4th Workshop on African Natural Language Processing.

[4] Chukwuneke, C., Ezeani, I., Rayson, P., &El-Haj, M. (2022, June). IgboBERT models: Building and training transformer models for the Igbo language. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 5114-5122).

[5] Emezue, C. C., Okoh, I., Mbonu, C., Chukwuneke, C., Lal, D., Ezeani, I., ... &Nmezi, O. (2024). The IgboAPI Dataset: Empowering Igbo Language Technologies through Multi-dialectal Enrichment. arXiv preprint arXiv:2405.00997.

[6] Harrando, I. (2022). Representation, information extraction, and summarization for automatic multimedia understanding (Doctoral dissertation, Sorbonne Université).

[7] Muppavarapu, V., Ramesh, G., Gyrard, A., &Noura, M. (2021). Knowledge extraction using semantic similarity of concepts from Web of Things knowledge bases. Data &Knowledge Engineering, 135, 101923.

[8] Chatterjee, S. (2023, January). Answering Topical Information Needs Using Neural Entity-Oriented Information Retrieval and Extraction. In ACM SIGIR Forum (Vol. 56, No. 2, pp. 1-2). New York, NY, USA: ACM.

[9] Inuwa-Dutse, I. (2025). Naijanlp: A survey of nigerian low-resource languages. arXiv preprint arXiv:2502.19784.

[10] Usip, P. U., Ijebu, F. F., Udo, I. J., &Ollawa, I. K. (2023). Text-Based Emergency Alert Framework for Under-Resourced Languages in Southern Nigeria. In Semantic AI in Knowledge Graphs (pp. 111-126). CRC Press.

[11] Ekle, O. A., &Das, B. (2025). Low-Resource Neural Machine Translation Using Recurrent Neural Networks and Transfer Learning: A Case Study on English-to-Igbo. arXiv preprint arXiv:2504.17252.

[12] Nie, B., Shao, Y., &Wang, Y. (2025). Improving distantly supervised named entity recognition by emphasizing uncertain examples. Pattern Analysis and Applications, 28(1), 13.

[13] Barrios González, E., Tovar Vidal, M., De Ita Luna, G., &Reyes-Ortiz, J. A. (2022, May). Extraction of Entities in Health Domain Documents Using Recurrent Neural Networks. In International Conference on Pattern Recognition and Artificial Intelligence (pp. 395-406). Cham: Springer International Publishing.

[14] Li, Q., Yao, N., Zhou, N., Zhao, J., &Zhang, Y. (2023). A joint entity and relation extraction model based on efficient sampling and explicit interaction. ACM Transactions on Intelligent Systems and Technology, 14(5), 1-18.

[15] Chaudhary, A. (2022). Automatic Extraction and Application of Language Descriptions for Under-Resourced Languages (Doctoral dissertation, Carnegie Mellon University).

[16] Tedeschi, S. (2025). Towards comprehensive and efficient information extraction across languages.

[17] Jembere, D. (2025). Breaking Barriers: Enhancing Universal Dependency Parsing for Amharic Advancing NLP for A Low-Resource Language.

[18] Jiang, X., Liang, Y., Chen, W., &Duan, N. (2022, June). Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 10840-10848).

[19] Jean, G. (2023). Cross-Lingual Transfer Learning for Low-Resource NLP Tasks: Leveraging Multilingual Pretrained Models.

[20] Kalyan, K. S., Rajasekharan, A., &Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542.

[21] Rigouts Terryn, A., Hoste, V., &Lefever, E. (2022). Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. Terminology, 28(1), 157-189.

[22] Jafari, A. (2022). Comparison Study Between Token Classification and Sequence Classification In Text Classification. arXiv preprint arXiv:2211.13899.

[23] Arora, S., Dalmia, S., Yan, B., Metze, F., Black, A. W., &Watanabe, S. (2022). Token-level sequence labeling for spoken language understanding using compositional end-to-end models. arXiv preprint arXiv:2210.15734.

[24] Chen, S., Huang, S., Pandey, S., Li, B., Gao, G. R., Zheng, L., ... &Liu, H. (2021, November). Et: re-thinking self-attention for transformer models on gpus. In Proceedings of the international conference for high performance computing, networking, storage and analysis (pp. 1-18).

[25] Zuo, R., &Mak, B. (2022). Local Context-aware Self-attention for Continuous Sign Language Recognition}}. Proc. Interspeech 2022, 4810-4814.

[26] Rahali, A., &Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. Ai, 4(1), 54-110.

[27] Luo, J., Wang, J., Cheng, N., Xiao, E., Xiao, J., Kucsko, G., ... &Li, J. (2021, July). Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[28] Lu, H., Wu, J., Ruan, Y., Qian, F., Meng, H., Gao, Y., &Xu, T. (2023). A multi-source transfer learning model based on LSTM and domain adaptation for building energy prediction. International Journal of Electrical Power &Energy Systems, 149, 109024.

[29] Lin, J., Ma, J., Zhu, J., &Liang, H. (2021). Deep domain adaptation for non-intrusive load monitoring based on a knowledge transfer learning network. IEEE Transactions on Smart Grid, 13(1), 280-292.

[30] Antonini, A., &Brooker, S. (2023, September). Name Links: an Aesthetic Discussion. In Proceedings of the 34th ACM Conference on Hypertext and Social Media (pp. 1-6).

[31] Bateman, J. A. (2022). A semiotic perspective on the ontology of documents and multimodal textuality. Using documents: A multidisciplinary approach to document theory, 147-198.

[32] Antia, B. E., &Mafofo, L. (2021). Text annotations: Examining evidence for a multisemiotic instinct and the intertextuality of the sign in a database of pristine self-directed communication. In Integrational linguistics and philosophy of language in the Global South (pp. 84-103). Routledge.