University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 8, Issue 1, 2024
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

State of the Art Approaches in Named Entity Recognition.

Catherine Dupe Omidiji*[1] , Emeka Ogbuju[2], Taiwo Abiodun[1], Joshua Jimba[1], Francisca Oladipo[3].

[1]Department of Computer Science, Federal University Lokoja, Nigeria.
[2] School of Computing, Miva Open University, Abuja. Nigeria.
[3]Thomas Adewumi University, Oko, Kwara State. Nigeria.

catherine.omidiji-msc@fulokoja.edu.ng; emeka.ogbuju@fulokoja.edu.ng; taiwo.abiodun-pg@fulokoja.edu.ng; Joshua.jimba@fulokoja.edu.ng; francisca.oladipo@fulokoja.edu.ng

Abstract: Named entity recognition (NER) is significant in extracting and categorizing entities from unstructured textual data, and it's a pivotal domain in Natural Language Processing (NLP). However, many researchers lack the appropriate method to conduct NER effectively. To address this issue, we conducted a methodical review, by sourcing related scientific papers from reputable scientific databases such as Scopus, IEEE Xplore, Science Direct and SpringerLink. Our study answered three research questions pertaining to the common approaches used for NER, commonly used algorithms for NER and how well have the algorithms have performed, and the state-of-the-art dataset commonly used for NER. The finding from the review showed a predominant adoption of machine learning, deep learning, hybrid model and rule-based approaches. Additionally, finding from this study shows the promising performance of the Conditional Random Field (CRF) and Bidirectional Long Short-Term Memory (BiLSTM) based models, particularly when hybridized. Despite the observed advancement of machine leaning and deep learning techniques in NER, this study find out some inconsistencies in the exiting studies, which includes reporting of dataset standards, and call for standardized practices. This study further proposed a novel NER framework based on the dataset nature, strategy in modeling, and application domain, highlighting areas that are underexplored like low resource language and domain-specific datasets. This study provides insights that can guide future research on NER, and proposes a method that can be used to enhance NER model's robustness, multilingual adaptability, and methodological transparency.

Keywords: Named entity recognition; machine learning; deep learning; rule-based; hybrid models; algorithms.

1. Introduction

In the rapidly evolving landscape of NLP and information retrieval, Named entity recognition (NER) hold an important role in series of application such as question answering, information extraction, sentiment analysis and many others [1]. NER system involves the identification and categorization of named entities within textual data, thereby facilitating the extraction of structured text. In the previous year's NER systems have has advanced significantly, with lot of researchers striving continuously to enhance their adaptively and performance to different languages and domain [2]. The evolvement of NER systems can be attributed to the proliferation of textual data in documents, over the internet, and other unstructured source. As software systems, where evaluation metrics are important for understanding software improvements [3], performance evaluation metrics also play an important role in advancing the effectiveness of NER models. Due to the continuous increase in data volume, there is a need for automated and more accurate NER systems. Realizing this, we delve into the historical development of NER and how it has adapted to address the specific needs and challenges of different eras.

In this paper, we present a comprehensive systematic literature review on the previous approaches utilized by different studies on NER systems. We focus on their methodologies, challenges, dataset and their relevance in evolving of NLP systems. NLP systems does not only focus on extracting information from text but also holds a substantial potential for real word application, such as automation of information retrieval, and the construction of knowledge gap. Most of these advancement is centered to machine learning, which has become an important tools to many field like health care, media etc. [4].

The aim of this study is to present an in-depth study of the state of the art techniques in the development of NER systems. This systematic literature review aims to provide a comprehensive reference for researchers, practitioners, and students in the field of NER, offering a detailed overview of the past, present and future of NER systems. By understanding the past and current state of NER system, readers can gain valuable insight into how the basic NLP task continue to shape the way information are extracted and make sense of unstructured textual data.

The subsequent sections of this paper are structured to provide a comprehensive exploration of the topic. We delve

into the methodology employed in our systematic literature review, detailing the steps we undertook to collect and analyze the relevant studies. Following that, we present the results of our review, showcasing the key findings and trends. Finally, the discussion section offers an in-depth analysis of the results, highlighting the significance and implication in the context of NER research and development.

2. Methodology

To conduct this review, we adopt a procedure performed by [5]. This method divided the review process into 3 stages, which includes; review planning, review conduction and review reporting

2.1 Review Planning

The first phase of this review is the planning phase, in this phase we ascertain the importance of the review. The purpose of carrying out this systematic literature review is to analyze and summarize the research on NER systems that were published from 2018 to 2023. This review's objectives are to summarize the prevailing methods used in NER, to identify the common algorithms for NER, to identify the state-of-the art dataset being used for NER and to identify gaps that exist in previous research.

After identifying the aim and the objectives of the research, we proceeded to formulating three research questions using the recommended guideline by [5]. Based on the recommended guideline, in other to formulate research question we must consider three view point of the study criteria which include population (higher institution), intervention (methods and algorithms for prediction), and outcome (successful prediction approaches). Hence, this study criterion has led to the formulation of the following research questions;

- RQ1: What are the common approaches used for NER?
- RQ2: What are the commonly used algorithms for NER and how well have they performed?
- RQ3: What are the state-of-the-art dataset commonly used for NER?.

2.2 Conducting the review

2.2.1 Data collection

To secure quality papers in this research, papers were extracted from four different databases, these databases include Scopus, IEEE Xplore, Science Direct and SpringerLink. A total of 294 papers were extracted from the mentioned scientific databases. However, after considering the papers based on the inclusion and the exclusion criteria, 28 papers were found relevant and were finally selected for the purpose of this review.

2.2.2 Screening process

A total of 294 papers were initially collected from the four academic database considered in this study, these papers were then pass through three screening process. Firstly, 6 duplicated papers were removed from the papers, and 288 papers were left for the second screening stage. During the second screening stage, 128 papers were removed based on the title and abstract that are not relevant to this study, leaving a total of 160 for the third screening stage. During the third screening stage, the papers were subjected to a full text review where 132 papers were screened and 28 papers were finally selected for this review. Each paper was independently assessed by two reviewers. Inter-rater agreement was calculated using Cohen's Kappa ($\kappa = 0.82$), indicating strong agreement. Disagreements were resolved through discussion. The PRISMA flow diagram for the reviews is presented in Table1.

The summary of relevant papers that were extracted in each database is tabulated in Table 1. This search was carried out in October, 2023.

Table 1: Initial number of papers extracted from each database

S/N	Database	Number of papers
1	Scopus	29
2	IEEE Xplore	92
3	Science Direct	76
4	SpringerLink	97
Total numbers of papers collected		**294**

Table2. Summary of final paper selection

S/N	Database	Number of papers
1	Scopus	3
2	IEEE Xplore	8
3	Science Direct	7
4	SpringerLink	10
Total numbers of papers collected		**28**

2.2.3 Search method

The search method that was used to extract relevant papers from the databases in Table 1 involves formulating the search string by using the following phrases; "Machine learning", "named entity recognition" "precognitive model", "data mining", and "information retrieval". Boolean operators like AND, and OR were used as conjunctions in the search strings.

2.2.4 Inclusion Criteria

- Papers published between 2018–2023
- Peer-reviewed journal or conference papers

- Papers that are focused on NER using machine learning or deep learning
- Papers written in English

2.2.5 Exclusion Criteria

- Papers published between 2018–2023
- Peer-reviewed journal or conference papers
- Papers that are focused on NER using machine learning or deep learning
- Papers written in English

2.2.6 Quality Assessment and Ranking

This study ensures quality and minimizes bias by subjecting each paper to a score of 0 to 5 point based on methodological clarity, dataset quality, relevance to the research questions, depth of evaluation, and Innovation. e included studies with a score 3 and above.

2.2.7 Normalization of Evaluation Metrics

Due to that fact that the performance metrics used by each study varies, performance metrics were not directly compared. Instead of this, we compared model performance within the domain (e.g., biomedical domain vs. general NER). Instead of absolute ranking, relative performance trends within domains and models were synthesized.

2.2.8 PRISMA flow diagram for the review's methodology

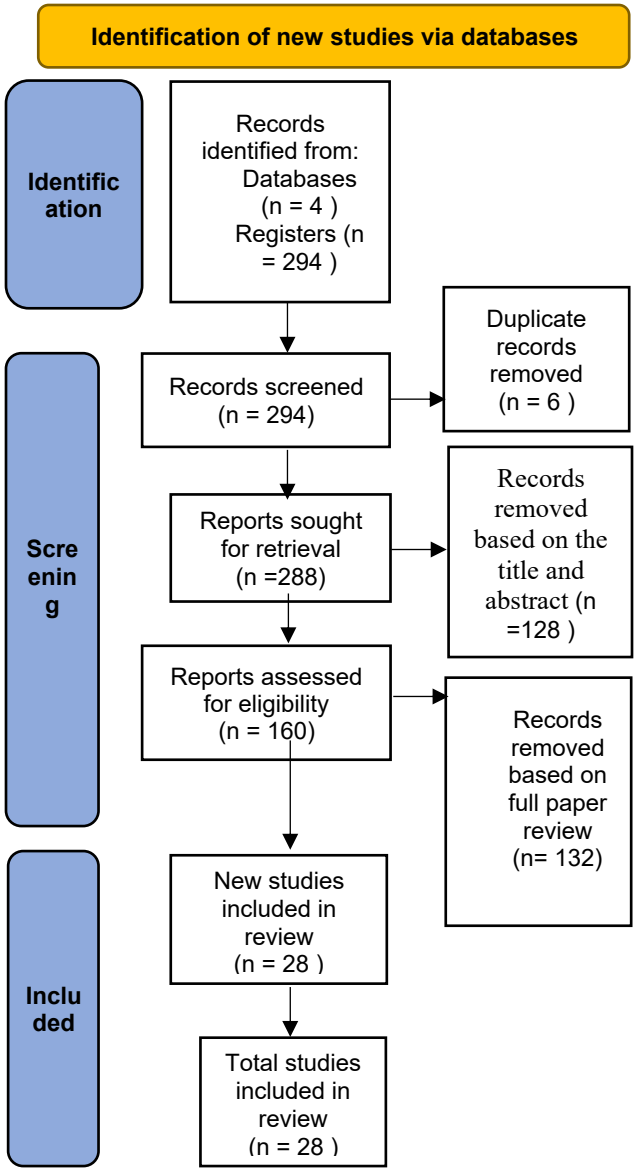

Figure 1. PRISMA flow diagram for the review's methodology

3. Review of selected papers

This section presents the details of the selected papers for this research, which includes the applied methods, dataset that has been used and the archived result. The aim is to provide a comprehensive review of these studies' objectives.

[6] adopted a deep learning method for NER in bidding document. The combination of the BiLSTM and The CRF was used to train the model. The dataset that was trained is a collection of 20,000 bidding documents. The proposed model was compared with other state-of-the-art algorithms like

convolution neural network (CNN), Long Short-Term Memory (LSTM), and BiLSTM. The proposed algorithm was able to outperform all other compared algorithm in terms of precision, recall, and f1-score, thus, achieving the highest precision, recall, and f1-score of 89.34%, 91.34%, and 90.24% respectively.

Different deep learning algorithms were used in developing a NER for Indonesian languages by [7]. The researchers focused on the use of a hybrid of BiLSTM and CNN architecture. The model focuses on the extraction of information from the language into four different class which includes organization, person, location and event. The corpus dataset was used for the building of the model, the Copus dataset consist of 4,139 sentences from news articles and some other articles about history of Indonesia. The result shows that the proposed model was able to outperform other compares algorithm and was able to achieve precision, recall and F1-score of 79.72%, 79.61% and 79.43% respectively.

A paper based on the development of a NER system for the purpose of detecting tumor morphology mentions in medical document was carried out by [8]. Machine learning algorithms such as CRF, BiLSTM and Bidirectional Encoder Representations from Transformer (BERT), and BiLSTM-CFR were used to train the model. The Chemist corpus was used for the building of the model, this dataset consists of information from 3,000 clinical cases which were stored in different files. From the result, the BiLSTM-CRF with pre-trained word embedding was able to perform best in terms of precision which is 82.8%, while the BERT obtained the best result in terms of recall and F1score which are 77.5% and 76.5% respectively

A study was with the aim to develop a NER was carried out by [9]. The study NER was developed to create tags that help fast information retrieval, query processing, and processing of data for tourism and travel. The authors of the paper made use of a domain specific knowledge based machine learning approach to build the model. The dataset that was used to develop the model is a manually curated Tourism and travel domain dataset from TripAdvisor and Wikipedia. The CRF algorithm was used to train the model. The result from the experiment shows that the model achieved a recall of 82%, precision of 85%, and accuracy of 82% and F-measure of 83%.

[10] used a deep learning architecture to build a Hindi text NER model. In order to boost the power of the BiLSTM which has been used by many other researchers, the authors proposed the de-noising auto-encoder (DEA) LSTM and conditioning LSTM. ICJNLP 2008 dataset which consist of 19822 sentences, 490368 total tokens in and 12 categories of entities which includes person, location, organization, brand, abbreviation, time, number, destination, measure title: person and title: object. The result shows that the DAE LSTM performed best in terms on F1 score, precision and recall with values of 72%, 81% and 66% respectively.

A clinical NER that can extract name entities from clinical narratives was developed by [11]. The study examined two deep learning algorithms which are recurrent neural network (RNN) and CNN, to extract name entities from clinical text. The algorithm was trained using the MIMIC II corpus and i2b2 2010 (training: 27837 entities, testing: 45,009 entities) dataset. The result of the experiment shows that the RNN model that was trained with word embedding achieve the best result in terms of F1 score which attain 85.94%.

[12] incorporated n-grams with BiLSTM and CRF to make a contextual long-short-term-memory network (CLSTM) for the purpose of building a biomedical NER system. The proposed model was compared with other state-of-the-art deep leaning algorithms such as BiLSTM, GRAM-CNN and BERT. Three corpora was used to train the model the corpora includes; National Center for Biotechnology Information (NCBI) containing 792 units, BioCreative II Gene Mention (GM) with 20,000 units, and the BioCreative V Chemicals diseases relationship (CDR) with 1500 units. From the result we observed that the all deep learning algorithms uses perform well. However, the proposed CLST was able to outperform all other compared algorithm in terms of precision, recall and F1score. The proposed model achieves the highest precision of 88.27% on CDR dataset, the highest recall of 86.67% in NCBI dataset and the highest F1score of 86.44% on CDR dataset.

[13] proposed an Arabic NER system that is based on the rule-based approach. The General Architecture for Text Engineering (GATE) was applied on the ANERcrop which consist 316 articles, including 150,286 tokens in the category of persons, organizations and location. The result of the experiment shows that the model was able to achieve F1score of 83% on person name entity, 89% on organization name entity and 92% on location name entity

[14] proposes the uses of character WC-CNN to replace the character WC-BiLSTM for Arabic NER purpose. The model was trained on the Automated Context Extraction (ACE) 2003 newswire and broadcast news (112,00 tokens), ANERcrop (150,000 tokens), and Twitter (81,000 tokens) datasets. The result shows that both model performed well on all datasets but have the best performance on the ACE2003 dataset. The WC-BiLSTM has the best F1 score of 94.92% on the ACE2003 dataset while the WC-CNN has the best F1Score of 91.47% on ACE2003 dataset.

A paper that introduces a rule-based NER method for classical Arabic document was published by [15]. The method that was proposed uses trigger word, gazetteers, grammatical rule, regular expression and blacklist. The CANER Corpus was used as dataset which contains over 7,000 Hadiths from Sahih Al-Bukhari book, and 72,108 named entity count. The result from the experiment shows that the proposed method was able to achieve 90.2% precision, 89.3% recall and 89.5% F-measures.

[16] presented a paper on the NER for five Nigerian languages which include Pidgin, Nigerian English, Igbo, Yoruba, and Hausa languages. The dataset that was used to train the model was locally curated by the researchers. The dataset was curated by collecting 50 different online articles

for the 5 languages. A hybrid approach consisting of rule-based and machine learning method was used to train the model. The model was train individually on each language and the trained on the combined languages. The result from the model shows that the hybrid model was able to achieve the highest f1-score, precision and recall of 67.13%, 70.29%, and 64.24% respectively on the Pidgin language dataset.

[17] proposed a NER in Yoruba text using a supervised machine learning approach. The model was trained using the CRF algorithm. The authors developed a Yoruba NER corpus by collecting religions based news in Yoruba language, which was used to train the model, the corpus contain 11, 617 tokens. The data set was divided into three features and the model was trained on each features. Three features that were used include the word-Internal features, word-external features and context features. The result from the experiment shows that the model achieved the best precision on the context feature which is 88.89%, and best recall and f1-score on the word-external features which are 69.70% and 77.97% respectively.

[18] proposes a NER methodology for English language. Natural language toolkit (NLTK) was used to perform the experiment and the method was compared to other state of the art methods. Dataset contain texts in English that were retrieved from different online platforms. The proposed method was evaluated and the result shows that the methods outperformed all other compared methods by achieving an accuracy of 94.75%.

[19] performed an experiment that compares the performance of the multilingual BERT algorithm on two low resource languages which are the Yoruba and the Twi languages. The dataset that was used consist of corpora collected by the Niger-Volta Language Technologies Institute, and it contains text from different sources, including online Bible, Yoruba language websites and Lagos-NWU conversation. The result shows that the model performed the best on the clean Yoruba dataset achieving the best accuracy of 60.9%.

A research with a goal to extract ADR related entities from Chinese ADER text was carried out by [20]. The researchers' proposed three different models, which includes CRF, BiLSTM-CRF and Lexical Features based BiLSTM-CFR (LF-BiLSTM-CRF). Chinese texts from ADER containing 147,451 entities were used to train the model. The result from the experiment shows that LF-BiLSTM-CRF our performed al other compared algorithm with the highest average F1-score of 94.35%.

[21] presented an approach to NER in the presence of scarce data. The proposed approach takes inspiration from question answering to recognize span an unseen domain. The pertained BERT was used to build the model. The researchers used some publicly available dataset which includes OntoNotes5.0, Conll2003, reviews from MIT restaurant and movies, ATIS and SNIPS. The result from the experiment shows the proposed method performed significantly well, especially when small amount of support examples is being used.

[22] conducted a study that proposed a conditioned-random-field based method that uses both language-independence and language-dependence features like context windows and POP respectively for NER. The researchers developed an Urdu NER dataset called the UNER-1 which consists of 58,633 tokens. The researches proposed a CRF algorithm to train the model. The propose model performed well achieving its highest precision, recall and F1-score of 88.21%, 84.05% and 86.68% respectively.

[23] presented a method that identifies fine grained multilingual named entities such as vehicles and musicalGRP. The research used the MultiConerV2 dataset which consist of 358,668 instances different languages such as English, French, Chinese, German, Hindi, Spain, Ukrainian, Swedish, Portuguese, Italian, Farsi and Bangla. XLM-RoBERTa was used as the baseline model, and other multilingual models that were used include, mLUKE, mDEBERTA, and mBERT.

[24] propose the use of FoodIE algorithm for recognizing food named entities. The researchers applied a rule-based approach which include POS tagging and semantic rule. The experiment was performed on food dataset that consist of 5 different classes which includes breakfast/lunch, snacks, drink, dinner, and dissert. The dataset was manually collected from different online source. The dataset consists of 200 different food items. The result of the experiment that was performed shoes that the proposed method achieved a precision of 97.8%, f1-score of 96% and recall of 94.3%, which make the proposed method better that the compare drNER method.

[25] developed a NER system that identifies chemical formulas, medical names and their relations. The method uses a hybrid of dictionary-based annotator and corpus-based disambiguation component. The model was trained on the CRAFT corpus which contains medical words in English language with entities such as organism, disease, molecular function, biological process, protein, cellular component, cell, line cell, chemical and sequence. The result shows that the NER mode achieved 71.4% f1-score.

[26] investigated a method that compared two different semi-supervised approaches which are expected maximization (EM) and label propagation (LP) to NER of Ethiopian news agency (ENA), with entities such as person, organization, location, money and date. The dataset that was used consist of text in Amharic language from the ENA which have of about 4700 sentences with approximately 83 word per sentence. The evaluation result shows that the EM achieved an F1 score of 61% and the LP achieved an f1-score of 79%, thus achieving the best result.

[27] propose a triggered NER system to address the problem of using large amount of data to get high accurate result. The researchers verified the effectiveness of entity triggers before creating a dual learning framework for low resource NE triggers and their corresponding entities. The

result of the experiment shows that the method that was proposed used 20% triggered entity annotated data and was able to achieve result comparable with the conventional methods which are trained by 50% - 80% data.

A dynamic NER entity-based approach under unconstrained tagging scheme was proposed by [28]. The researchers recognize some commonly used tagging schemes and then proposed two novel unconstrained scheme in order to remove the constrain. A dynamic mechanism that dynamically address the input by the interaction between the output label and the input text. The dataset that was used to train the mode contain are gotten from different sources and contains texts in languages such as English, Spanish, Dutch and German. The result of the experiment shows that the proposed method can perform well on different language, by achieving accuracies more that 80%.

[29] proposed a novel approach for information extraction from different scientific source. This approach was proposed to address the problem of low accuracies. The novel ontological approach uses the word sense disambiguation (WSD) and the fuzzy rule-based (FRB). The dataset contains information from scientific databases such as IEEE, Springer ACM, and Elsevier. The result shows that the proposed method was able to achieve a significant accuracy of 89.14% and F1-score of 89%.

A deep learning approach for extracting Covid-19 symptom from social media was proposed by [30]. A novel deep learning model that integrates both semantic and syntactic analysis for symptom extraction and classification was proposed by the researchers. The researchers utilized the Multilayer Perception (MLP). The dataset that was used in this experiment is the Twitter chatter dataset which contained Covid19 related tweets on Twitter that was posted from March 15 2020, to May 09, 2020. The result of the experiment shows that the proposed model outperformed all other compared models and achieved recall, precision and F1-score of 92%, 90% and 91% respectively on the training data.

A methodology for NER in Portuguese language was proposed by [31]. The proposed method utilized the LSTM algorithm architecture for the purpose of NER. The dataset set that was used has categories such as dates, organizations, locations, specific codes, monetary values etc. The result from the experiment shows that the proposed method was able to achieve maximum precision of 83.38% on Portuguese language.

[32] built an Amharic NER system with the use of Transformer Based Recognizer. The deep learning algorithms used to build the model includes the RoBERTa, and the bidirectional LSTM coupled with CRF. The researchers introduced a new Amharic named entity dataset that was used to train the model. The result from the experiment shows that the RoBERTa performed best and was able to achieve an F1-score of 93%.

An attempt to address underrepresentation of Africa languages in NLP research by creating an NER for African languages was carried out by [33]. The researchers made used of different algorithms such as CNN-BiLSTM-CFR, mBERT, XLM-R, and MeanE-BiLSTM. The African dataset that was used is the MasakhaNER dataset which contain African languages like Amharic, Igbo, Hausa, Luganda, Kinyarwanda, Nigerian Pidgin, Luo, Swahili, Yoruba and Wolof languages from different sources. The result from the experiment shows that the XLM-R has the highest average accuracy of 78.81% on all languages. However, the compared algorithms performed well by achieving average accuracies more that 70% on all languages.

Table 3: Taxonomy of NER Approaches Based on Methodology and domain of Application

NER Approaches	Subcategories by Domain & Language Specificity	Examples
Traditional ML	General purpose NER Domain specific NER Multilingual NER	CRF, SVM used in biomedical HMM in Hindi or Yoruba
Deep Learning	General purpose NER Domain specific NER Multilingual NER	BiLSTM, BERT XLM-R
Hybrid	General purpose NER Domain specific NER Multilingual NER	CNN-BiLSTM-CRF Rule-based + ML for medical or legal documents Hybrid approach for Pidgin
Rule-Based Systems	General purpose NER Domain specific NER Multilingual NER	Regex for English Grammar rules for Arabic texts Gazetteers for Yoruba

In other to provide a better structured synthesis of the methods used in NER, this study proposed a taxonomy that classifies NER approaches into two parts which are the method used and the domain and language specificity. Based on the literature that were reviewed in this study, we have observed that NER approaches can be classified into four different methodological types which are the traditional machine learning, deep learning, hybrids and the rule-based methods. Additionally, we observed that the methodological categories can be sub classified into general purpose, domain specific and multilingual NER approaches.

The proposed taxonomy presents in Table 3 serves as a synthesis framework for organizing and evaluating prior research. For example, BiLSTM-CRF and BERT which are deep learning based approaches have been dominant across general and domain-specific NER tasks, particularly in biomedical and clinical domains. While the Rule-based methods are still being utilized in specific contexts, like classical Arabic and other low-resource African languages due to their interpretability and ease of implementation.

However, we still observed some gaps in this landscape which includes underutilization of hybrids model for multilingual settings, despite their potential to combine rule-based insights with the adaptability of neural architectures. Additionally, while the deep learning approaches have shown promising result in lot of domain, cross-lingual transfer and domain adaptation still remain underexplored, especially in low resource African languages. This taxonomy highlights these gaps and provides a structured lens for guiding future research in NER.

3.1 Evolution Timeline of NER Methods

Early applications of NER methods were dominated by the rule-based and the traditional machine learning approaches like the CRF and HMM, especially in low resource languages like Yoruba [17] and domain specific areas like tourism [9]. Between 2015 and 2020 the deep learning approach rises, algorithms like notably BiLSTM and BiLSTM-CRF were mainly used in biomedical [12], clinical [11], and multilingual settings [7, 10]. Recently, researches have shin focus n the transformer models which has now emerge a the state-of-the-art tools since 2019, due to its outstanding performance in both general-purpose and low-resource NER tasks [33, 32]. Now recent studies are exploring the hybrid approach to improve model's performance in low-resource contexts [27].

4. Result

4.1 Common Approaches Used for

This section presents the approaches that are commonly used in NER research based on the reviews, the commonly used approaches by researchers for NER are found to be machine learning, deep learning, hybrid, and the rule-based approaches. Table 4 presents the common approaches used in NER. As shown in Table3, the deep learning approach is mostly used for NER as the largest part of the reviewed paper which is 39.29% actually utilized the deep leaning approach, a significant percentage of the reviewed papers of 32.14% utilized the hybrid models approach, 21.43% utilized the machine learning approach while only 25% make use of the rule-based approach.

Table4: Common approaches used for NER

Papers	Methods	Frequency
[8, 17, 9, 20, 21, 26]	Machine Learning	21.43%
[8, 10, 11, 14, 19, 21, 23, 30, 31, 32, 33]	Deep Learning	39.29%
[6, 7, 8, 12, 16, 20, 25, 32, 33]	Hybrid models	32.14%
[13, 15, 18, 24, 27, 28, 29]	Rule-Base	25%

The common approaches used in NER presented in Table 4 shows that four approaches are mainly utilized by previous studies. These approaches are machine learning, deep learning, hybrid models, and rule-based approaches. We observed that machine learning models were used in 21.43% of the reviewed studies. The deep learning approaches were utilized by 39.29% of the reviewed papers. Hybrid approaches which combine the abilities of two or more approaches to improve NER performance and adaptability over a wide range of datasets and domains were utilized y 32.14% of the evaluated publications. While only 25% of the reviewed papers utilized the rule-based approaches, which extract items from text using predetermined language rules and patterns. The rule-based approaches are less common than the machine learning and the deep learning approaches, but are viable choice, especially in situations where domain-specific knowledge and linguistic patterns are critical for accurate entity extraction.

The deep learning models approaches are commonly utilized because of their effectiveness in the development of NER systems. Also hybrid approaches are commonly used in NER systems because of their ability to combine the strengths of different approaches to improve NER systems. These approaches help researchers to get high accuracies in NER tasks, and extend NER ability in different applications like information extraction, knowledge discovery, and natural language understanding. Researchers can gain insights into the changing landscape of NER methodologies and make right decisions on the techniques to select and implement based on some specific research objectives by examining the efficacy of the common approaches used by previous studies.

4.2 Commonly used Algorithms for NER and how well they have performed

Details of the commonly used algorithms used in NER and their performance are presented in Table 5, highlighting the methods that have been used by previous studies to extract entities from textual data.

Table 5: Commonly used Algorithms for NER and how well they have performed.

Papers	Algorithms	Performance
[6]	BiLSTM-CRF	precision, recall, and f1-score of 89.34%, 91.34%, and 90.24% respectively.
[7]	BiLSTM-CNN	precision, recall and F1-score of 79.72%, 79.61% and 79.43% respectively.
[8]	CRF, BiLSTM BERT and BiLSTM-CRF	BiLSTM-CFR with best precision of 82.8%, and BERT with the best recall and F1score of 77.5% and 76.5% respectively.
[10]	(DEA) LSTM, and conditioning LSTM	DAE LSTM performed best in terms on F1 score, precision and recall with values of 72%, 81% and 66% respectively.

[9]	CFR	a recall of 82%, precision of 85%, accuracy of 82% and F-measure of 83%
[11]	RNN and CNN	RNN achieve the best result in terms of F1 score which attain 85.94%.
[12]	CLSTM	Best precision of 88.27% on CDR dataset, the best recall of 86.67% in NCBI dataset and the best F1score of 86.44% on CDR dataset
[13]	GATE	the model achieve F1score of 83% on person name entity, 89% on organization name entity and 92% on location name entity
[14]	WC-BiLSTM and WC-CNN	The WC-BiLSTM has the best F1 score of 94.92% on the ACE2003 dataset while the WC-CNN has the best F1Score of 91.47% on ACE2003 dataset.
[15]	trigger word, gazetteers, grammatical rule, regular expression and blacklist	the proposed method was able to achieve 90.2% precision, 89.3% recall and 89.5% F-measures.
[17]	CRF	the model achieved the best precision on the context feature which is 88.89%, and best recall and f1-score on the word-external features which are 69.70% and 77.97% respectively.
[19]	multilingual BERT	the model performed the best on the clean Yoruba dataset achieving the best accuracy of 60.9%
[20]	CRF, BiLSTM-CRF and LF-BiLSTM-CRF	LF-BiLSTM-CRF our performed al other compared algorithm with the highest average F1-score of 94.35%.
[21]	BERT	Precision above 80%
[21]	CRF	The propose model performed well achieving its highest precision, recall and F1-score of 88.21%, 84.05% and 86.68% respectively.
[24]	FoodIE	precision of 97.8%, f1-score of 96% and recall of 94.3%
[25]	hybrid of dictionary-based annotator and corpus-based disambiguation component	71.4% f1-score
[26]	EM and LP	EM achieved an F1 score of 61% and the LP achieved an f1-score of 79%
[29]	WSD with FRB	89.14% and F1-score of 89%
[30]	MLP	recall, precision and F1-score of 92%, 90% and 91% respectively on the training data
[31]	LSTM	maximum precision of 83.38% on Portuguese language.
[32]	RoBERTa, and the bidirectional LSTM coupled with CRF	RoBERTa performed best and was able to achieve an F1-score of 93%
[33]	CNN-BiLSTM-CFR, mBERT, XLM-R, and MeanE-BiLSTM	XLM-R has the highest average accuracy of 78.81% on all languages. However, the compared algorithms performed well by achieving average accuracies more that 70% on all languages.

The commonly used algorithms in performing NER tasks and their performance across various research have been summarized in Table 5. This comprehensive summary presents detail of the different approaches that have been used for NER tasks, shedding light on the state of NER researches. These algorithms include CRF, BiLSTM, BiLSTM-CRF, BiLSTM-CNN, BERT, multilingual BERT, LSTM, RNN, CNN, CLSTM, GATE, WC-BiLSTM, WC-CNN, LF-BiLSTM-CRF, hybrid methods, EM, LP, WSD with FRB, MLP, and RoBERTa. These algorithms have produces promising results when utilized for NER tasks. Although some of these algorithms have outperform others, and proven to be more effective in performing NER tasks. For example, CRF and BiLSTM algorithms are very effective in studies conducted by [22, 8, 17, 14, 9], as they were able to achieve high accuracy, precision, recall, and F1-scores. Additionally, some studies utilized the hybrid methods

Hybrid techniques that combine the strengths of two different approaches was also utilized in some studies. For example, [20] utilized the LF-BiLSTM-CRF to achieve F1-score of 94.35%, and [6] utilized BiLSTM-CRF, to achieve precision, recall, and F1-scores of 89.34%, 91.34%, and 90.24%, respectively. This result shows the effectiveness of hybridizing different algorithms to improve models performance.

4.3 State-of-the-art Dataset Commonly used for NER

An overview of the state-of-the-art dataset that has been utilized in the review papers on NER is presented in Table 6. The compilation serves as a comprehensive reference, shedding lights on the diverse datasets that have played a pivotal in advance the field of NER.

Table 6: State-of-the-art dataset commonly used for NER

Paper	Dataset	Description
[7]	news articles and other articles about history of	4,139 sentences

	Indonesia, written in Indonesia language.	
[8]	Chemist dataset consisting of information from 3,000 clinical cases	Not stated
[10]	ICJNLP 2008 dataset (Hindi text)	19822 sentences, 490368 total tokens
[11]	MIMIC II corpus and i2b2 2010	72846 total tokens
[12]	National Center for Biotechnology Information (NCBI), BioCreative II Gene Mention (GM) and the BioCreative V Chemicals diseases relationship (CDR)	NCBI= 792 tokens GM = 20,000 tokens CDR = 1,500 tokens
[13]	ANERcrop (Arabic text)	316 articles, including 150,286 tokens
[14]	Automated Context Extraction 2003 newswire (ACE2003) and broadcast news, ANERcrop, and Twitter dataset	ACE2003 = 112,00 tokens ANERcrop = 150,000 tokens Twitter = 81,000 tokens
[15]	CANERCorpus	77,000 articles and 72,108 tokens
[20]	Chinese text from ADER	147,451 entities
[21]	OntoNotes5.0, Conll2003, reviews from MIT restaurant and movies, ATIS and SNIPS	Not stated
[21]	UNER-1 Urdu dataset	58,633 tokens
[23]	MultiConerV2 dataset	358,668 instances
[25]	CRAFT corpus	790,000 tokens, 67 articles
[26]	Manualy collected corpus from Amharic Ethiopian news agency	4,700 sentences with approximately 83 word per sentence
[33]	MasakhaNER dataset	Not stated

A summary of the state of the art dataset used in NER research is presented in Table 6. The summary table shows significant differences in the way different studies report the dataset used. While some of these studies show details of the dataset, with thorough description, including number of articles, sentences, and tokens, other study provides little information about the used dataset, omitting important details. For example, studies carried out by [21, 8], and [26] utilized manually curated dataset which consists of texts collected from different online sources such as Twitter, and news agencies, but these studies does not report key indicators. On the other hands, [10, 25] include detailed information about the dataset used. These divers' ways of presenting dataset makes it difficult to compare studies, and also present the important of standardizing reporting procedures throughout the NER research community. The adoption of uniform dataset repot could help in the improvement of NER research consistency. Additional the comprehensive summary presented in Table 5 covers different languages and domain, demonstrating the global scope of NER research. However, we observed a clear imbalance in dataset representation, with some languages and domains overrepresented while other are underrepresented. Efforts should be focused on the development of more balanced dataset that includes both high resource and low resource languages. The adoption of uniform and standard report of dataset should be encourage by the NER research community, as it will help in the development of more robust models capable of handling varied languages and domains. This will help to advance NER globally, allowing for more effective information extraction and knowledge discovery from text data.

5. Discussion

In this review, we were able to address three research questions to shed light to the landscape of NER. Our analysis from the review a series of approaches used in NER which include machine learning, deep learning, hybrid models and rule-based approach. Deep learning constitutes 39.29% of the review papers being the highest approach being used for NER showcase its significance in advancing NER methodologies. Hybrid models and rule-based approaches adopted considerably, underscoring the divers' strategies employed by researchers in tackling NER challenges.

The second research question concerning the commonly used algorithms for NER and how well they have performed, our review uncovered a plethora of algorithms, among which the CRF and the BiLSTM stood out with remarkable performance. However, the combination of CRF and BiLSTM exhibited superior outcomes in terms of F1-score, precision, recall, and accuracies. These findings show the potential synergy achievable by integrating different algorithms, emphasizing the important of thoughtful algorithmic selection in optimizing NER performance.

Finally, our analysis on the state-of-the-art dataset for NER elucidated a rich tapestry of linguistic diversity, drawing from different language in different part of the world. However, we observed inconsistencies in reporting standards, especially regarding dataset that were manually curated. Thus, we recommended standardized reporting practices to enhance clarity and facilitate comparative analyses.

5.1 Critical Analysis and Novel Classification of NER Approaches

This study provides deep insight on the state of NER research by proposing a classification framework based on the data type, techniques in modeling, and application domain. This framework will allow researchers to make significant improvement on NER solutions.

1. Data type: previous studies mostly rely on semi-structured text such as news, clinical, and reports. However, little studies have addressed unstructured and noisy dataset like multilingual data. Also previous studies overrepresented high resource

language like English and French, but under represent low resource language

2. Modeling Techniques: This study proposed This study proposed four different approaches based on the nature of the dataset. The proposed techniques are the traditional ML (e.g., CRF, HMM) which is suitable for small, and labeled dataset, the deep learning (eg BiLSTM, CNN) due to its automatic feature extraction nature, Hybrid approach (e.g., BiLSTM-CRF, mBERT + Rule-based) due to its effectiveness in noisy and multilingual environment, and the transformer-based approaches (e.g., BERT, RoBERTa) which is underutilized for low resource languages
3. Healthcare, Finance, Legal, Social Media: Previous studies mostly focused on healthcare and news data. Focus should be expanded to underrepresented sectors like agriculture, indigenous governance documents, and e-commerce reviews in African countries.

5.2 Critical Synthesis

From the synthesis of the reviewed papers, we observed that hybrid models performs better than the standalone models in term of accuracy and F1-score, and the transformer models are underutilized in low resource language particularly African languages, due to the limitations in computational and annotation. We also observed that the reporting of standard dataset and benchmarking protocols are lacking, which limits comparability and reproducibility. Additionally, most existing studies does not report error analysis or biasness in model.

6 Conclusion

This review paper has provided an understanding of the current state of NER methodologies. The prevalence of the machine learning, deep learning, hybrid models and rule-based approaches highlight the evolving significance of advance techniques in NER research. The identification of CRF and BiLSTM as standout algorithms, particularly when combined, emphasizes the potential for synergistic effect in optimizing NER performance. However, the analysis on the state-of-the-art dataset revealed inconsistencies in reporting standards, urging the need for a standardized report practice to enhance clarity and facilitate comparative analyses. As diversity in languages continue to play important role in NER, future efforts should prioritize the creation of balance dataset that contains lots of divers languages. As NER landscape expands, insights garnered from this review paper will serve as a guide to researchers in the field of NER, highlighting the importance of algorithm selection and combination, and the ongoing pursuit of linguistic inclusive. This synthesis does not only contribute to the current body of knowledge but also lays the foundation for future endeavors aimed at refining methodologies, fostering applicability of cross-linguistic and most importantly advancing the efficiency of NER.

References

[1] Pejic-Bach, M., Bertoncel, T., Krst, Z., & Mesko, M. (2020). Text mining of industry 4.0 job advertisements. *Int J. Inf. Management, 50*(1), 416-431.

[2] Naseer, S., Ghafoor, M. M., Alvi, K. S., Kiran, A., Rahman, S. U., & Murtaza, G. (2021). Named entity recognition (NER) in NLP techniques, tools accuracy and performance. *Pakistan Journal of Multidisciplinary Research, 2*(2), 293-308.

[3] Fatima, S., Fatima, Z., Hayat, M. A., Shahab, M. H., Meraj, M. K., Ibrahim, R. M., & Muneeb, S. M. (2022). Impact of software metrics on software quality using McCall quality model: In-depth analysis. *University of Sindh Journal of Information and Communication Technology (USJICT)*, *6*(2), 37–46.

[4] Hyder, H., Ali, K. H., Aziz, A., & Iram, L. (2024). Evaluating diabetes detection methods: A multilinear regression approach vs. other machine learning classifiers. *University of Sindh Journal of Information and Communication Technology (USJICT)*, *7*(2), 47–56.

[5] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Linkman, S., & Bailey, J. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology, 15*(1), 7-15.

[6] Ji, Y., Tong, C., Liang, J., Yang, X., Zhao, Z., & Wang, X. (2019). A deep learning method for named entity bidding document. *Journal of physics, 1168*, 1-11.

[7] Gunawan, W., Suhartono, D., Purnomo, F., & Ongko, A. (2018). Named-entity recognition for Indonesian Language using Bidirectional LSTM-CNNs. *3rd International Conference on Computer Science and Computational Intelligence 2018.135*, pp. 425-432. Jakarta, Indonesia: elsevier.

[8] Romero, G. d., & Segura-Bedmar, I. (2020). Exploring deep learning for named entity recognition of tumor morphology mentions. *Proceeding of the Iberian Languages Evauation Forum*, *2664*, pp. 1-16. Madrid, Spain.

[9] Vijay, J., & Rajeswari, S. (2018). A machine learning apprach to named entity recgnition for the travel and turism domain. *Asian Jurnal f Infrmatin Technology, 15*(21), 4309-4317.

[10] Shah, B., & Kopparapu, S. K. (2019). A deep learning approachfor Hindi named entity recognition. *arXiv*, 1-7.

[11] Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2018). Clinical named entity recognition using deep learning models. *AMIA annual Symposium proceedings.2017*, pp. 1812-1819. AMIA Symposium.

[12] Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics, 20*(735), 1-11.

[13] Elsherif, H. M., Alomari, K. M., Shaalan, K., & Alhamad, A. Q. (2019). Arabic rule-based named entity recognition system using GATE. *15th International conference on Machine Learning and Data Mining, MLDM 2019*, (pp. 1-15). New York, USA.

[14] Khalifa, M., & Shaalan, K. (2019). Character convlutions for Arabic named entity recognition with long short-term memory networks. *Computer Speech and Language, 58*(2019), 335-346.

[15] Salah, R., Mukred, M., Zakaria, Q. B., Ahmed, R., & Sari, H. (2022). A new rule-based approach for classical Arabic in natural language processing. *Journal of Mathematics, 2022*, 1-20.

[16] Oyewusi, F. W., Adekanmbi, O., Okoh, I., Onuigwe, V., Salami, I. M., Osakuade, O., . . . Musa, A. U. (2021). NaijaNER: comprehensive named entity recognition for 5Nigerian languages. *Computation and Language*, 1-5.

[17] Ayogu, I., Adetunbi, A. O., & Ayogu, B. A. (2019). A first step towards the development of Yoruba named entity recognition system. *International Journal of Computer Application, 182*(41), 1-4.

[18] Jain, R., Sharma, A., Mishra, G. S., & Nand, P. (2020). Named entity recognition in English text. *Journal of Physics, 1712*(1), 1-6.

[19] Alabi, J. O., Amponsah-Kaakyire, K., Adelani, D. I., & Espana-Bonet, C. (2020). Massive vs. curated embeddings for low-resourced languages: a case study of Yoruba and Twi. *proceedings of the 12th*

Conference on Language Resources and Evaluation (LREC 2020) (pp. 2754-2762). Marceille : Eutopean anguage Resoources Association.

[20] Chen, Y., Zhou, C., Li, T., Wu, H., Zhao, X., Ye, K., & Liao, J. (2019). Nemed entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. *Journal of Biomedical Informatics, 96*(2019), 1-8.

[21] Ziyadi, M., Sun, Y., Goswami, A., Huang, J., & Weizhu, C. (2020). Example-based named entity recognition. *Computing Research Respository , abs-2008-10570*(2020).

[22] Khan, W., Shahzad, K., Amjad, T., Banjar, A., & Fasihuddin, H. (2022). Named entity recognition using conditional random field. *Appied Science, 12*(13), 1-18.

[23] Fetahu, B., Kar, S., Chen, Z., Rokhlenko, O., & Malmasi, S. (2023). SemEvsl-2023 Task2:FIne-grained multilingual named entity recognition (MultiCoNER2). *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2032)* (pp. 2247-2265). Toronto, Canada: Association for Computational LLinguistics.

[24] Popovski, G., Kochev, S., Seijak, B. K., & Eftimov, T. (2019). FoodIE: A rule-based named entity recognition method for food information extraction. *The INternational Conference on Pattern Recognition Applications and Methods*, *1*, pp. 915-922. Prague, Czech Republic.

[25] Furrer, L., Jancso, A., Colic, N., & Rinaldi, F. (2019). OGER++: Hybrid multi-type entity recognition. *Journal of Cheminformatics, 11*(1), 1-10.

[26] Hirpassa, S., & Lehal, G. S. (2020). Named entity recognition: a semi-supervised learning approach. *International Journal of Information Technology, 13*(3), 1659-1665.

[27] Zhong, M., Liu, G., Xiong, J., & Zuo, J. (2022). DualNER: A triggerbased dual learning framework for low-resource named entity recognition. *IEEE Intelligent System, 37*(4), 79-87.

[28] Zhao, F., Gui, X., Huang, Y., Jin, H., & Yang, L. T. (2022). Dynamic entity-based named entity recognition under unconstrained tagging schemes. *IEEE Transaction on Big Data, 8*(4), 1059-1074.

[29] Zaman, G., Mahdin, H., Hussain, K., & Rahman, A. (2021). An ontological framework for information extraction from diverse scientific sources. *IEEE Access, 9*, 42111-42121.

[30] Luo, X., Gandhi, P., Storey, S., & Huang, K. (2022). A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media. *IEEE J Biomed Health Inform, 26*(4), 1737–1748.

[31] Silva, R. A., Silva, L., Dutra, M. L., & Araujo, G. M. (2021). An improved NER methodology to the Portuguese language. *Mobile Networks and Applications, 26*(1), 319-325.

[32] Jibril, E. C., & Tantug, C. (2023). ANEC: An Amharic named entity corpus and tranformer based recognizer. *IEEE Access, 11*(1), 15799-15815.

[33] Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., . . . Mayhew, S. (2021). MasakhaNER: named entity recognition for African languages. *Transactions of the Association for Computational Linguistics, 9*(5), 1116-1131.