Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 03 March-2025, Page No.- 4314-4325

DOI: 10.47191/etj/v10i03.38, I.F. – 8.482

State-Of-The-Art Named Entity Recognition and Related Extraction: A Review

Firdaws Rizgar Tato[1], Ibrahim Mahmood Ibrahim[2]

[1]Akre University for Applied Sciences Technical College of Informatics Department of Information Technology

[2]Akre University for Applied Sciences, Technical College of Informatics Department Computer Network and Information Security

ABSTRACT: Named Entity Recognition (NER) has evolved significantly as a key component in the field of Natural Language Processing (NLP). This review paper encapsulates the progress and trends in NER by exploring state-of-the-art techniques and methodologies employed across various domains. It highlights the shift from traditional rule-based models to advanced machine learning approaches, including deep learning and transformers, which have markedly enhanced the performance of NER systems. Particular emphasis is given to the adaptation of NER for specific needs such as biomedical information extraction, cybersecurity, and multilingual entity recognition, reflecting the growing complexity and diversity of application fields. Recent advances demonstrate the integration of sophisticated technologies like graph attention networks and multimodal frameworks, which leverage both contextual and syntactic features to address the challenges of polysemy and entity disambiguation. The review also discusses the crucial role of domain-specific adaptations, the importance of large, annotated datasets, and ongoing efforts to mitigate limitations related to data scarcity in low-resource languages. This comprehensive overview not only sheds light on the technological advancements but also sets the stage for future explorations aimed at further refining the accuracy and applicability of NER systems across more diverse and challenging datasets.

KEYWORDS: Transformer Models Multimodal Integration Domain-Specific Adaptations Entity Disambiguation Data Annotation

1 INTRODUCTION

Named Entity Recognition (NER) and related extraction are pivotal areas of research in the field of Natural Language Processing (NLP). NER involves identifying named entities in text and categorizing them into predefined classes such as person, organization, and location. This task is fundamental for various downstream applications, including information retrieval, question answering, and entity linking [1]. The significance of NER is highlighted by its role as a key building block in information extraction systems [2], and its crucial applications in specialized domains like cyber threat intelligence [3], biomedical information extraction [4][5], and combating human trafficking [6].

Recent advancements in NER research have been driven by the exploration of advanced techniques such as transformer-based models, which have shown improved accuracy and efficiency due to their ability to capture long-range dependencies and contextual information [7][8]. Transfer learning frameworks, such as T2NER, based on transformers have demonstrated promising results in enhancing NER accuracy [9]. Additionally, approaches like BERT-based transfer learning, self-augmentation, and meta reweighting have been proposed to further boost the robustness and performance of NER systems [10][11][12].

The evolution of NER systems has also involved the integration of sophisticated technologies like graph attention

Networks, which have been leveraged to fuse contextual and syntactic features, particularly in domains such as biomedical named entity recognition (BioNER) [13][14]. This specialization is crucial for identifying entities like genes, proteins, diseases, and drugs, with methods such as label re-correction and knowledge distillation being explored to enhance performance [15].

Moreover, significant progress has been made in developing specialized corpora and recognizers for specific languages and domains, reflecting the efforts to cater to diverse linguistic requirements [16][17]. Research in low-resource scenarios has introduced novel strategies like pre-training models to enhance NER efficiency in environments with limited data [18][19].

Additionally, tackling the challenges of ambiguity and polysemy in named entity recognition remains a critical area for further research. Entities often have multiple meanings or are used in different contexts, which can confuse traditional NER systems. Advanced disambiguation techniques, which leverage contextual clues and world knowledge, are required to discern the correct meanings of entities in varying contexts [20][21]. Furthermore, the issue of fine-grained entity classification presents another layer of

complexity. Moving beyond broad categories like person, location, and organization to more nuanced sub-categories could enable NER systems to provide more detailed insights, beneficial for applications such as precision marketing and in-depth content analysis [22][33]. Addressing these nuanced challenges not only requires improvements in algorithmic approaches but also in the quality and diversity of training datasets. Efforts to generate and curate high-quality, annotated datasets that reflect a wide range of languages, dialects, and domain-specific jargon are essential for training more robust and accurate NER systems [23].the continuous innovation in NER and related extraction techniques has significantly advanced NLP capabilities. By leveraging cutting-edge methodologies, incorporating domain-specific knowledge, and addressing challenges such as entity overlap and discontinuity, researchers aim to push the boundaries of NER performance and applicability across various domains and languages.

The remainder of this paper is structured as follows: Section 2 details the review analysis, exploring the evolution from rule-based to advanced deep learning models in NER and their domain-specific adaptations. Section 3 synthesizes key conclusions, highlighting the implications for fields requiring precise entity recognition. Section 4 discusses the broader ramifications of these advancements and challenges such as the dependency on annotated datasets. Finally, Section 5 concludes by summarizing the study's contributions to Natural Language Processing and suggesting directions for future research.

2 THEORETICAL BACKGROUND

State-of-the-art named entity recognition (NER) is at the forefront of advancements in natural language processing (NLP), aiming to accurately identify and classify entities within text into predefined categories such as persons, organizations, locations, and others. This domain has seen rapid evolution, particularly with the adoption of advanced machine learning techniques and deep learning models.

2.1 Current Trends in NER

Recent developments in NER are characterized by the use of deep learning architectures, especially Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach), which have significantly enhanced the accuracy of entity recognition across diverse datasets and languages [24][25]. These models leverage large-scale language modeling and fine-tuning on specific NER tasks to achieve state-of-the- art results.

2.2 Deep Learning Innovations

Transformers utilize self-attention mechanisms that inherently capture contextual relationships between words in a sentence, regardless of their position [26]. This ability has led to breakthroughs in not just NER but also in related tasks such as sentiment analysis, question answering, and more. The introduction of models such as ALBERT (A Lite BERT) and DeBERTa (Decoding-enhanced BERT with Disentangled Attention) further refine the efficiency and effectiveness of these approaches, offering improvements in both speed and performance [27][28].

2.3 Integration with Other NLP Tasks

Modern NER systems are increasingly integrated with other NLP tasks to enhance overall text understanding. Tasks like relation extraction, coreference resolution, and entity linking are often performed in conjunction with NER to provide a more comprehensive analysis of text data [28] . This integration facilitates a deeper understanding of text semantics, supporting advanced applications in information extraction and knowledge graph construction.

2.4 Challenges and Future Directions

Despite advancements, NER systems still face challenges, such as dealing with polysemy and homonymy in entity names, domain-specific variations, and low-resource languages [29][30]. Future research is expected to focus on cross-lingual NER, few-shot learning, and the integration of unsupervised learning techniques to address these challenges.

3 RELATED WORK

Bin Jia (2020) develop an advanced Named Entity Recognition (NER) approach for Chinese electronic medical records (CEMRs), targeting malignant tumor entities. Utilizing a combination of BiLSTM-CRF models for sequence handling and CNNs for feature extraction, they implement a collaborative framework to address the complexities inherent in medical data. They enhance their model's capability through sentence-level transfer learning from non-target scene datasets, compensating for the sparse distribution of entities in the target dataset. This approach achieves an impressive F1-score of 87.60% on a competitive dataset, showcasing the effectiveness of their multi- model architecture. However, the model's reliance on multiple complex systems and extensive pre-training could hinder its broader application. To increase adaptability and efficiency, the study suggests streamlining the architecture by refining the integration of models and potentially reducing the number of networks involved [31].

Ahmed, Khurshid (2024) introduce a novel deep learning-based model, RoBERTa-BiGRU-CRF, for joint extraction of cyber entities and relations from cyber threat intelligence, improving cyber defense. Employing a relation-matching technique, the model efficiently handles complex CTI data. It surpasses previous models with an impressive 7% increase in F1-score, achieving 0.86. Despite its strengths, the model's effectiveness is limited to specific CTI data types and distributions. To enhance its generalizability, the authors suggest integrating adaptive algorithms to better accommodate diverse cyber threat environments [32].

Caiyu Wang (2020) enhance Chinese medical named entity recognition (NER) by integrating a multi-granularity semantic dictionary with a multimodal tree to address the complexities of Chinese clinical texts. Their approach uses

LSTM and CRF models, optimizing the fusion of character and word information to improve semantic depth and accuracy. The algorithm specifically targets the reduction of errors from Chinese word segmentation, significantly boosting the model's performance across various datasets. Despite its effectiveness, the model heavily relies on a detailed semantic dictionary, limiting its broader adaptability. To overcome this, expansion of the dictionary and incorporation of adaptive learning techniques are suggested to enhance versatility across medical subdomains [34].

Cong Sun (2021) innovate in biomedical named entity recognition (BioNER) by integrating BioBERT into a machine reading comprehension (MRC) framework, moving away from traditional sequence labeling. This approach uses queries to contextualize entity recognition tasks, making use of BioBERT's deep understanding of biomedical context. The technique abolishes the need for sequence-based decoding, using BioBERT to directly predict entity boundaries from text. This method demonstrated state-of-the-art performance, achieving high F1-scores across multiple datasets, such as 92.92% on BC4CHEMD. The reliance on extensive domain-specific data for optimal performance poses a limitation. The model could be improved by diversifying training materials and refining query strategies to enhance general applicability and robustness [35].

Hermenegildo Fabregat (2023) investigate the application of negation-based transfer learning to enhance Named Entity Recognition (NER) and Relation Extraction (RE) in biomedical texts, focusing on disabilities and rare diseases. They employ Bi-LSTM networks and CRFs, leveraging pretrained weights from a negation detection model to boost NER and RE model accuracy. This technique significantly improves model performance, showing up to a 13% increase in F-Measure for NER tasks and 2% for RE tasks in both English and Spanish. The model's effectiveness is limited by its dependency on the training data's specificity, suggesting expansion of training datasets and further optimization of the negation detection component as potential enhancements [36].

Rodrigo Juez-Hernandez (2023) develop AGORA, a system designed for the anonymization, extraction, and mapping of sensitive information from documents, facilitating secure data sharing between public institutions and research centers. AGORA employs advanced Named Entity Recognition (NER) models alongside anonymization algorithms to identify and anonymize personal data within documents, ensuring data privacy while maintaining utility for research. The system integrates geoparsing techniques to further process and visualize the data. Successfully tested in police and medical domains, AGORA proves effective in enhancing data security and usability for analytical purposes. However, its performance is contingent upon the quality and format of the input documents. To broaden its application, it is suggested that AGORA be enhanced to handle a wider variety of document formats and improve robustness across different real-world scenarios [37].

Nasi Jofchea (2022) present a method for Named Entity Recognition (NER) in pharmaceutical texts using transfer learning with models like spaCy, AllenNLP, BERT, and BioBERT. They focus on enhancing the accuracy of detecting specific entities such as Pharmaceutical Organizations and Drugs. The methodology involves fine-tuning pre-trained models on accurately labeled, domain-specific datasets, which improves the recognition accuracy significantly over generic models. They achieved impressive F1 scores of 96.14% for known entities and 95.14% for unseen entities in the training data, showcasing the method's effectiveness. The approach is designed for integration into systems that provide actionable medical information to patients. However, the model's performance is heavily dependent on the training data's domain specificity. To improve, they suggest expanding the datasets to cover broader contexts and incorporating more general training methods [38].

Xu Jiang (2024) introduce APIE, an advanced information extraction module that integrates the best aspects of pipeline and joint approaches for Named Entity Recognition (NER) and Relation Extraction (RE). APIE utilizes separate encoders for each task to avoid feature confusion and incorporates multi-level attention for NER and local context pooling for RE. This design allows APIE to handle complex document-level data effectively, improving accuracy and reasoning capabilities significantly compared to traditional methods. The model demonstrates superior performance across multiple benchmarks, offering a robust solution for comprehensive information extraction challenges. However, Thomas Saout, (2024) provide an overview of automated data extraction from invoices, crucial for business accounting and tax purposes. They explore various automation techniques involving Optical Character Recognition (OCR), natural language processing (NLP), and machine learning to streamline the extraction and processing of invoice data. The paper emphasizes advanced machine learning models that tackle the complexity and variability of invoice formats. These technologies are shown to significantly enhance the accuracy and efficiency of invoice processing by reducing manual entry errors and processing time. However, their effectiveness is limited by the dependency on high-quality digital inputs and structured formats. To address these limitations, the authors suggest developing more robust OCR and machine learning algorithms capable of handling diverse invoice types and quality levels [48].

Mariana Dias (2020) assess Named Entity Recognition (NER) for sensitive data in Portuguese, focusing on compliance with GDPR. They utilize a hybrid approach combining rule-based, lexicon-based, and machine learning techniques, specifically Conditional Random Fields (CRF), Random Forest, and Bidirectional-LSTM models. These methods aim to enhance the accuracy and efficiency of identifying and classifying sensitive data within unstructured texts. The Bidirectional-LSTM model demonstrated superior

performance, achieving an F1-score of 83.01%, showcasing its efficacy in sensitive data recognition. However, the limited availability of extensive Portuguese language datasets restricts model training and overall accuracy. The paper suggests expanding the training datasets with a broader range of Portuguese texts to improve the robustness and applicability of NER systems [49].

Yu Wang (2020) introduce ERNIE-Joint, an advanced model for Chinese Named Entity Recognition (NER) that integrates ERNIE's enhanced representation capabilities with joint training for NER and text classification. By utilizing both token-level and sentence- level features, ERNIE-Joint optimizes entity recognition and classification simultaneously without needing additional annotations. This approach leverages ERNIE's knowledge masking strategy, which surpasses BERT's character-level strategy by focusing on entity and phrase levels, leading to superior recognition accuracy. The model achieves state-of-the-art results on the MSRA-NER and Weibo datasets, outperforming both BERT and standard ERNIE. However, its effectiveness is primarily confined to Chinese text, which may limit broader application. Expanding training to include multilingual datasets and adapting the model for various text types could potentially broaden its applicability and improve its utility [50].

Miguel A. (2021) delve into the integration of parsing techniques in Named Entity Recognition (NER), highlighting the benefits of using syntactic structures to enhance entity identification and delimitation. They advocate for treating parsing as a sequence labeling task, which simplifies integrating deep syntactic insights into NER systems without complex modifications. This approach, they argue, significantly improves NER accuracy by leveraging sentence structures, traditionally overlooked in basic sequence labeling models. However, the computational demands of parsing pose practical limitations for widespread application in NER tasks. They propose increasing the efficiency of parsing algorithms and simplifying their integration as potential solutions to these challenges, aiming to broaden the application of syntactic analysis in NER systems [51].

Priyankar Bose (2021) explore advancements in Named Entity Recognition (NER) and Relationship Extraction (RE) within clinical texts, aimed at improving healthcare information extraction. Their review covers a spectrum from traditional machine learning to modern deep learning techniques, noting a shift towards automatic feature learning and the integration of complex data types. Particularly, deep learning models that utilize embeddings and neural networks are emphasized for their proficiency in managing unstructured clinical data. These models have shown improved accuracy and practical applicability in healthcare settings. However, their effectiveness is significantly dependent on the availability of high-quality, annotated clinical datasets, which are often limited and costly to develop. The paper suggests that adopting semi-supervised and unsupervised learning methods could reduce dependence on extensive annotated data, potentially broadening the utility and efficiency of NER and RE systems in clinical contexts [52].

László Nemes (2021) investigate the integration of sentiment analysis with information extraction and named entity recognition to analyze public sentiment on social media during the COVID-19 pandemic. They employ a mix of RNN and BERT models, supplemented by NLP tools like NLTK and TextBlob, to deeply analyze tweet content. Their approach enhances traditional sentiment analysis by categorizing tweets and extracting entities, providing a nuanced understanding of public opinions. This methodological integration offers vital insights beneficial across disciplines such as linguistics and psychology. However, the effectiveness of their analysis is constrained by the dependence on high-quality labeled data. To overcome this limitation, they suggest expanding the dataset variety and incorporating unsupervised learning techniques to enhance the models' robustness and the accuracy of sentiment detection [53].

Fang, Li, (2023) introduce a multi-task learning-based Chinese Named Entity Recognition model, MTL-BERT, designed to simultaneously process entity boundaries and types, enhancing accuracy and learning efficiency. This model leverages BERT for foundational pre-training, complemented by task-specific encoding and decoding layers that handle distinct subtasks of boundary and type annotation. MTL-BERT significantly outperforms traditional single-task models, achieving higher F1 scores on various Chinese datasets including Weibo NER, MSRA, and OntoNote4.0. However, its complexity and reliance on specific data types could restrict broader application. To improve, the study suggests simplifying the multi-task framework and expanding training across diverse datasets to boost the model's versatility and applicability [54].

Liu, (2023) enhance Named Entity Recognition (NER) for social media by incorporating data augmentation techniques into their model, addressing the unique challenges of informal and unstructured text prevalent on these platforms. Their approach leverages BERT for semantic extraction combined with a Bi-LSTM-CRF framework and a self-attention layer to refine the recognition process. By employing methods like synonym replacement and semantic transformation, the model significantly increases the diversity and quality of training data. This leads to state-of-the-art performance on datasets such as WNUT16, WNUT17, and Onto Notes 5.0, showing marked improvements in entity recognition accuracy. However, the model's dependency on data augmentation techniques may limit its ability to fully capture the nuances of social media language. To enhance the model's robustness, further development of more sophisticated and context-aware data augmentation strategies is suggested [55].

Sawicki (2024) investigate the application of Named Entity Recognition (NER) and Graph Networks to uncover common

interests across thematic sub-fora on Reddit. Their approach combines NER to identify entities within subreddit posts with graph networks that analyze relational dynamics based on these entities. By linking entities through cross-posts across various subreddits, they create graph models that reveal the interconnections and shared themes among Reddit communities. This methodology has successfully identified new relationships and commonalities across subreddits, providing deeper insights into user behavior and community dynamics. However, their reliance on cross-posts, which are not universally frequent or representative, may limit the findings' generalizability. To improve, the researchers propose expanding data sources to include a broader range of subreddit interactions and employing predictive models to enhance the depth and accuracy of their analysis [56].

He, Wang, (2024) develop a Visual Clue Guidance and Consistency Matching Framework (GMF) for improving Multimodal Named Entity Recognition (MNER). This innovative approach integrates BERT for text processing and ResNet50 for image features, enriched with visual clues and cross-scale attention mechanisms. Their framework uniquely incorporates visual information using a consistency matching module based on multimodal contrastive learning, aiming to synchronize text and image modalities effectively. Notably, the model demonstrates superior performance on Twitter datasets from 2015 and 2017, achieving F1 scores of 75.81% and 87.11%, respectively. However, its complex requirements may limit deployment in environments with limited resources. To enhance usability, they could streamline the model to reduce computational demands and improve scalability [57].

Yi, Jiang,(2023) develop RDF-CRF, a multimodal ensemble learning model for cybersecurity Named Entity Recognition (NER) that effectively tackles complex cybersecurity texts. This model combines rule-based methods, a known-entity dictionary, and Conditional Random Fields (CRF) to enhance entity recognition accuracy. The approach starts with identifying potential entities using rule- based and dictionary-based methods, followed by CRF which refines these identifications using four innovative feature templates. This methodological synergy allows RDF-CRF to outperform existing models, achieving superior precision and recall on several cybersecurity datasets. However, its efficacy heavily relies on the comprehensive coverage of rule sets and dictionaries. Enhancing the model could involve continual updates to these resources and incorporating adaptable machine learning techniques to better recognize evolving entity variations in cybersecurity contexts [58].

Chih-Ming Tsai's 2023 study introduces the NER-SA model, which integrates Named Entity Recognition (NER) and Natural Language Processing (NLP) to detect fake news through stylometric analysis across different domains. The model leverages NLP to analyze textual content and NER to scrutinize the relationships between named entities, enhancing the detection of fabricated content. Employing a mathematical model optimized by simulated annealing, NER-SA significantly improves detection accuracy and F1 scores compared to existing methods, particularly on the LIAR and FakeNews AMT datasets. However, its effectiveness depends heavily on the completeness and accuracy of the named entity databases used. Enhancements recommended include updating these databases regularly and incorporating more dynamic entity recognition methods to maintain and improve detection capabilities [59].

Jennifer D'Souza (2024) develops the ORKG Agri-NER service, aimed at enhancing FAIR scholarly contributions in agriculture by employing advanced Named Entity Recognition (NER) integrated with the Open Research Knowledge Graph (ORKG) platform. This service utilizes a range of state-of-the-art neural architectures and transformer models to effectively identify and classify agricultural entities from scholarly texts. A multi-step entity resolution process aligns these entities with the AGROVOC ontology, improving the precision of entity classification. ORKG Agri-NER substantially enhances the machine-actionability of agricultural research data, promoting better data sharing and reuse. However, its effectiveness is limited by the current scope of the AGROVOC ontology. To overcome this, the study recommends expanding and regularly updating the ontology and employing more dynamic learning models to accommodate new agricultural concepts and terms [60].

Table 1- Summary of Recent Advances in Named Entity Recognition (NER) Techniques

Name/Year	Challenge	Algorithm/Approach	Research Focus/Area	Advantage	Disadvantage
Bin Jia, 2020	Chinese electronic medical records (CEMRs)	BiLSTM-CRF, CNNs, transfer learning	Medical NER	High F1-score, addresses data sparsity	Complex systems, extensive pre-training
Ahmed Khurshid, 2024	Cyber threat intelligence (CTI)	RoBERTa-BiGRU-CRF, relation-matching	Cyber NER and relations	7% increase in F1-score, efficiently handles CTI data	Limited to specific CTI types

Caiyu Wang, 2020	Chinese clinical texts	LSTM, CRF, multi-granularity semantic dictionary	Medical NER	Improves semantic depth and accuracy, reduces errors	Heavy reliance on detailed semantic dictionary
Cong Sun, 2021	Biomedical texts	BioBERT, MRC framework	Biomedical NER	Abolishes sequence-based decoding, state-of-the-art performance	Reliance on domain-specific data
Hermenegildo Fabregat, 2023	Biomedical texts on disabilities and rare diseases	Bi-LSTM, CRFs, negation-based transfer learning	Biomedical NER and RE	Significant improvement in model performance	Dependency on specific training data
Rodrigo Juez-Hernandez, 2023	Document anonymization and extraction	NER models, anonymization algorithms, geoparsing	Document security	Enhances data security and usability	Performance dependent on document quality
Nasi Jofchea, 2022	Pharmaceutical texts	spaCy, AllenNLP, BERT, BioBERT, transfer learning	Pharmaceutical NER	High F1 scores for known and unseen entities	Performance reliant on domain-specific training data
Xu Jiang, 2024	Information extraction	APIE, separate encoders, multi-level attention, local context pooling	NER and RE	Superior performance, handles complex data effectively	Dependence on specific encoder configurations
Hu Zhang, 2023	Legal documents	MRC framework, BERT-based model	Judicial NER	Enhances identification of nested entities	Primarily optimized for legal texts
Darshana Dash 2024	Bio-NER	BERT, ELMO, Bi-LSTM, CRF	Biomedical texts	Improves entity recognition	Dependent on high-quality pre-trained embeddings
Long Ding 2024	Few-shot NER	Encoder interventions, causal inference	NER with minimal data	Reduces bias, improves learning	Limited scalability and adaptability
S. Rizou 2023	Multilingual conversational agent	BiLSTM, CRF, pre-trained language models	Administrative services	Simplifies architecture, robust functionality	Limited to closed-domain applications
Asra Jehangir 2023	NER	CNNs, RNNs, LSTMs, Transformers	Various datasets	Enhances NER systems	Limited applicability in less-resourced languages
Yinxia Lou 2020	NER in low-resource domains	Graph Attention, BiLSTM-CRF, domain-specific dictionaries	Biomedical texts	Reduces dependency on large datasets	Contingent on dictionary quality
Gang Yang 2020	NER	Residual BiLSTM, CRF	NER	Boosts performance significantly	Specialized blocks may not be effective elsewhere

Gorjan Popovski 2020	NER for food-related information	Rule-based, corpus-based, deep neural networks	Food-related texts	High accuracy in food entity extraction	Dependency on potentially incomplete food ontologies
Thomas Saout 2024	Automated data extraction	OCR, NLP, machine learning	Invoices	Enhances accuracy and efficiency	Limited by input quality and structure
Mariana Dias 2020	NER for sensitive data	CRF, Random Forest, Bi-LSTM	GDPR compliance	High efficacy in sensitive data recognition	Restricted by dataset availability
Yu Wang 2020	Chinese NER	ERNIE-Joint (integration of ERNIE with NER and text classification)	Chinese texts	Optimizes recognition and classification simultaneously	Limited to Chinese text
Miguel A. Alonso et al. 2021	NER with parsing integration	Parsing as sequence labeling	NER	Enhances entity identification with syntactic structures	Computational demands of parsing
Priyankar Bose et al. 2021	NER and RE in clinical texts	Traditional ML to deep learning, embeddings, neural networks	Healthcare information extraction	Manages unstructured clinical data effectively	Dependent on high-quality annotated datasets
László Nemes 2021	Sentiment analysis with NER	RNN, BERT, NLTK, TextBlob	Social media sentiment analysis	Enhances sentiment analysis by integrating NER	Dependent on high-quality labeled data
Fang, 2023	Multi-task Chinese NER	MTL-BERT with encoding and decoding layers for entity boundaries and types	Chinese NER	Enhances accuracy and learning efficiency	Complexity and reliance on specific data types
Liu 2023	NER for social media	BERT, Bi-LSTM-CRF, self-attention, data augmentation techniques	Social media	Increases training data diversity and quality	Dependency on data augmentation techniques
Sawicki, 2024	NER and Graph Networks for Reddit analysis	NER with graph networks	Reddit community dynamics	Reveals interconnections and shared themes among subreddits	Reliance on cross-posts may limit generalizability
He, Wang, 2024	Multimodal NER	BERT, ResNet50, visual clue guidance, consistency matching framework	Multimodal entity recognition	Superior performance with visual clues and cross-scale attention	Complex requirements may limit deployment
Yi, Jiang 2023	Cybersecurity NER	RDF-CRF with rule-based methods, known-entity dictionary, CRF	Cybersecurity	Superior precision and recall on cybersecurity datasets	Efficacy depends on comprehensive rule sets and dictionaries

Chih-Ming Tsai 2023	NER for fake news detection	NER-SA model with NLP and simulated annealing	Fake news detection	Improves detection accuracy and F1 scores	Depends heavily on completeness of named entity databases
Jennifer D'Souza 2024	Agricultural NER integrated with ORKG	Advanced neural architectures, transformer models, multi-step entity resolution	Agricultural research	Enhances machine-actionability of agricultural research data	Limited by the scope of AGROVOC ontology

4 DISCUSSION

The field of Named Entity Recognition (NER) has seen significant technological advances driven by the adoption of machine learning models, particularly deep learning and transformers. These innovations have expanded NER's efficacy and applicability across diverse domains such as cybersecurity, biomedical information extraction, and multilingual entity recognition.

Recent progress includes the use of transformer-based models like BERT and RoBERTa, which enhance accuracy through advanced contextual understanding. Additionally, sophisticated methodologies such as graph attention networks and multimodal integration have been utilized to tackle challenges in polysemy and entity disambiguation. These approaches help in understanding the nuanced differences and contextual uses of language, crucial for domains where precision is paramount.

However, despite these advancements, several challenges persist. The dependency on large, annotated datasets is a significant hurdle, especially for low-resource languages where such data is scarce. Moreover, the complexity of entities and the need for domain-specific adaptations require continuous refinements in models and methodologies.

Future directions in NER research may focus on enhancing the robustness of these systems against the backdrop of evolving language use and domain-specific needs. This includes leveraging unsupervised and semi-supervised learning to reduce the reliance on extensive annotated corpora, thus making NER tools more adaptable and easier to deploy across various languages and specialized fields.

.

5 EXTRACT STATISTICS

The Figure 1 provides a clear view of the distribution of various algorithms and approaches used in Named Entity Recognition (NER) research. Notably, Conditional Random Fields (CRF) and BERT (Bidirectional Encoder Representations from Transformers) are the most frequently mentioned, indicating their widespread use due to CRF's effectiveness in context and dependencies in sequence prediction, and BERT's ability to capture deep contextual relationships. Bi-LSTM (Bidirectional Long Short-Term Memory) and its combination with CRF also appear significant, reflecting the integration of neural networks with graphical models to enhance accuracy. Emerging technologies like the Machine Reading Comprehension (MRC) Framework and domain-specific adaptations such as sBioBERT for biomedical texts highlight the specialized use of established models. The presence of transfer learning emphasizes its role in leveraging pre-trained models to improve performance on specialized datasets. Additionally, a variety of neural architectures and specific adaptations like RNN (Recurrent Neural Networks) underscore the diversity in NER approaches, with the category "All others each mentioned" indicating ongoing experimentation with novel algorithms alongside established ones. This graph effectively underscores the current technological trends and preferences within the NER research community

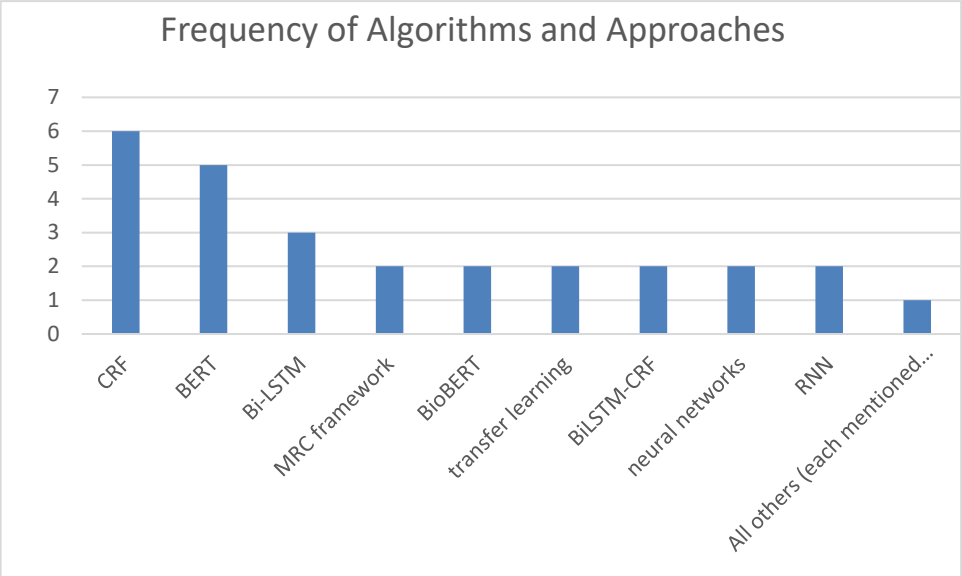

Figure 1: Frequency of Algorithms and Approaches

The histogram illustrates the frequency distribution of research focus areas, revealing that the majority of topics appear only once, indicating a diverse range of studies. A few areas, such as "Medical NER," "Biomedical texts," and "NER," appear twice, showing repeated research interest in these fields. The tall bar at frequency **1** represents unique topics, while the shorter bar at frequency **2** highlights the limited overlap in research focus. This distribution suggests that while certain topics receive more attention, the field remains highly diverse, with many studies exploring distinct challenges in Named Entity Recognition (NER) and related areas.

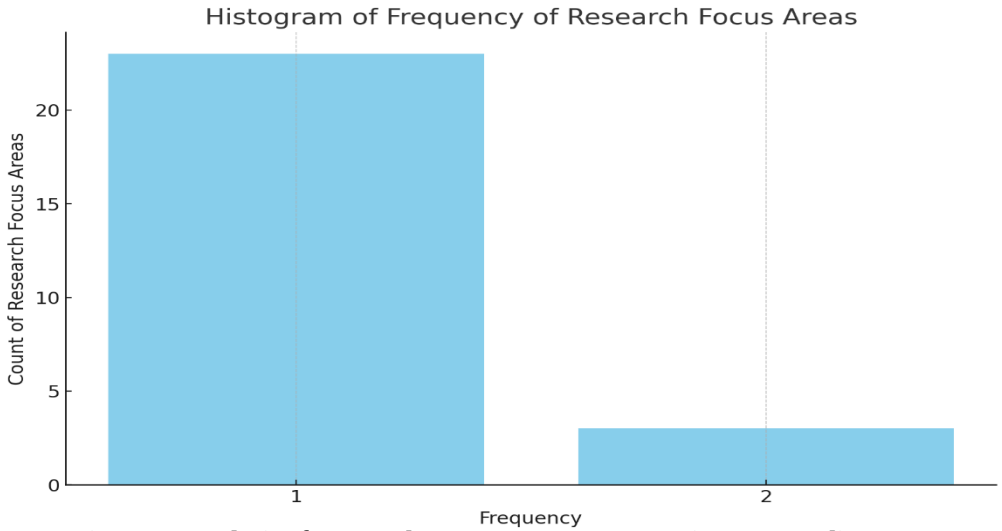

Figure 2: Analysis of Research Focus Area Frequency in NER Studie

5 CONCLUSION

This review of Named Entity Recognition (NER) encapsulates a broad spectrum of advancements and trends that have shaped the field into a pivotal component of modern Natural Language Processing (NLP). As we have examined, the evolution from rule-based systems to sophisticated machine learning frameworks, particularly deep learning and transformer-based models, has significantly enhanced the performance and applicability of NER systems. These models have been crucial in addressing complex NER tasks across diverse domains such as biomedical, cybersecurity, and multilingual contexts, offering increased accuracy and robustness.

The integration of technologies like graph attention networks and multimodal frameworks has notably improved the handling of syntactic and contextual nuances, essential for overcoming challenges like polysemy and entity ambiguity. Moreover, domain-specific adaptations have enabled the tailoring of NER systems to the intricate needs of specialized fields, demonstrating substantial gains in precision and utility.

However, despite these advancements, several challenges remain. The dependency on extensive, high-quality annotated datasets continues to be a significant barrier, particularly in low-resource languages and specialized domains. This underscores the necessity for continued efforts in dataset

creation and the exploration of semi-supervised and unsupervised learning techniques to alleviate the reliance on annotated data.

Looking forward, the field of NER is poised for further innovation. Future research will likely explore more advanced integration of unsupervised learning models, expand the capabilities of NER systems to handle increasingly complex and nested entities, and enhance the adaptability of these systems to dynamic and real-world applications. As NER continues to evolve, its impact across various sectors—ranging from healthcare to finance and beyond—promises to be profound, driving forward the capabilities of artificial intelligence in understanding and processing human language.

REFERENCES

1. S. Amin and G. ¨ Unter Neumann, "T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition." [Online]. Available: https://github.com/thuml/
2. A. D. P. Ariyanto, D. Purwitasari, and C. Fatichah, "A Systematic Review on Semantic Role Labeling for Information Extraction in Low-Resource Data," *IEEE Access*, vol. 12, pp. 57917–57946, 2024, doi: 10.1109/ACCESS.2024.3392370.
3. H. Lughbi, M. Mars, and K. Almotairi, "CybAttT: A Dataset of Cyberattack News Tweets for Enhanced Threat Intelligence," *Data (Basel)*, vol. 9, no. 3, p. 39, Feb. 2024, doi: 10.3390/data9030039.
4. H. Zhou, Z. Liu, C. Lang, Y. Xu, Y. Lin, and J. Hou, "Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation," *BMC Bioinformatics*, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s12859-021-04200-w.
5. D. Vithanage, P. Yu, L. Wang, and C. Deng, "Contextual Word Embedding for Biomedical Knowledge Extraction: a Rapid Review and Case Study," *J Healthc Inform Res*, vol. 8, no. 1, pp. 158–179, Mar. 2024, doi: 10.1007/s41666-023-00157-y.
6. Amina Catherine Ijiga, Enoch Joseph Aboi, Idoko Peter Idoko, Lawrence Anebi Enyejo, and Micheal Olumubo Odeyemi, "Collaborative innovations in Artificial Intelligence (AI): Partnering with leading U.S. tech firms to combat human trafficking," *Global Journal of Engineering and Technology Advances*, vol. 18, no. 3, pp. 106–123, Mar. 2024, doi: 10.30574/gjeta.2024.18.3.0046.
7. A. Rahali and M. A. Akhloufi, "End-to-End Transformer-Based Models in Textual-Based NLP," *AI (Switzerland)*, vol. 4, no. 1. Multidisciplinary Digital Publishing Institute (MDPI), pp. 54–110, Mar. 01, 2023. doi: 10.3390/ai4010004.
8. M. B. Shishehgarkhaneh, R. C. Moehler, Y. Fang, A. A. Hijazi, and H. Aboutorab, "Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia," *IEEE Access*, vol. 12, pp. 41829–41851, 2024, doi: 10.1109/ACCESS.2024.3377232.
9. S. Chen, Y. Pei, Z. Ke, and W. Silamu, "Low-resource named entity recognition via the pre-training model," *Symmetry (Basel)*, vol. 13, no. 5, May 2021, doi: 10.3390/sym13050786.
10. M. H. Syed and S. T. Chung, "Menuner: Domain-adapted bert based ner approach for a domain with limited dataset and its application to food menu domain," *Applied Sciences (Switzerland)*, vol. 11, no. 13, Jul. 2021, doi: 10.3390/app11136007.
11. A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, and N. Dragoni, "BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling," *Applied Sciences (Switzerland)*, vol. 12, no. 3, Feb. 2022, doi: 10.3390/app12030976.
12. R. Anam *et al.*, "A deep learning approach for Named Entity Recognition in Urdu language," *PLoS One*, vol. 19, no. 3 March, Mar. 2024, doi: 10.1371/journal.pone.0300725.
13. Y. Tian, W. Shen, Y. Song, F. Xia, M. He, and K. Li, "Improving biomedical named entity recognition with syntactic information," *BMC Bioinformatics*, vol. 21, no. 1, Dec. 2020, doi: 10.1186/s12859-020-03834-6.
14. X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, "BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-05051-9.
15. Y. Zhong and S. D. Goodfellow, "Domain-specific language models pre-trained on construction management systems corpora," *Automation in Construction*, vol. 160. Elsevier B.V., Apr. 01, 2024. doi: 10.1016/j.autcon.2024.105316.
16. Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named Entity Recognition and Relation Extraction: State-of-The-Art," *ACM Comput Surv*, vol. 54, no. 1, Apr. 2021, doi: 10.1145/3445965.
17. H. Alamro, T. Gojobori, M. Essack, and X. Gao, "BioBBC: a multi-feature model that enhances the detection of biomedical entities," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58334-x.
18. W. Gao, Y. Li, X. Guan, S. Chen, and S. Zhao, "Research on Named Entity Recognition Based on Multi-Task Learning and Biaffine Mechanism," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/2687615.
19. Y. Jiang, F. Jin, M. Chen, G. Liu, and Y. Yuan, "Cross-domain NER in Data-poor Scenarios for

Human Mobility Knowledge," 2023, doi: 10.21203/rs.3.rs-3152699/v1.
20. W. Bouarroudj, Z. Boufaida, and L. Bellatreche, "Named entity disambiguation in short texts over knowledge graphs," *Knowl Inf Syst*, vol. 64, no. 2, pp. 325–351, Feb. 2022, doi: 10.1007/s10115-021-01642-9.
21. A. Hur, N. Janjua, and M. Ahmed, "Unifying context with labeled property graph: A pipeline-based system for comprehensive text representation in NLP," *Expert Syst Appl*, vol. 239, Apr. 2024, doi: 10.1016/j.eswa.2023.122269.
22. W. Li, J. Liu, Y. Gao, X. Zhang, and J. Gu, "Chinese Fine-Grained Named Entity Recognition Based on BILTAR and GlobalPointer Modules," *Applied Sciences*, vol. 13, no. 23, p. 12845, Nov. 2023, doi: 10.3390/app132312845.
23. Y. Zhang and G. Xiao, "Named Entity Recognition Datasets: A Classification Framework," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1. Springer Science and Business Media B.V., Dec. 01, 2024. doi: 10.1007/s44196-024-00456-1.
24. J. Lee and J. A. Shin, "Decoding BERT's Internal Processing of Garden-Path Structures through Attention Maps*," *Korean Journal of English Language and Linguistics*, vol. 23, pp. 461–481, 2023, doi: 10.15738/kjell.23..202306.461.
25. C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.
26. B. S. Al-Smadi, "DeBERTa-BiLSTM: A multi-label classification model of Arabic medical questions using pre-trained models and deep learning," *Comput Biol Med*, vol. 170, Mar. 2024, doi: 10.1016/j.compbiomed.2024.107921.
27. E. Kotei and R. Thirunavukarasu, "A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning," *Information (Switzerland)*, vol. 14, no. 3. MDPI, Mar. 01, 2023. doi: 10.3390/info14030187.
28. M. S. I. Malik, A. Nazarova, M. M. Jamjoom, and D. I. Ignatov, "Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, Sep. 2023, doi: 10.1016/j.jksuci.2023.101736.
29. A. H. Oliaee, S. Das, J. Liu, and M. A. Rahman, "Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types," *Natural Language Processing Journal*, vol. 3, p. 100007, Jun. 2023, doi: 10.1016/j.nlp.2023.100007.
30. S. Harrer, "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine," *eBioMedicine*, vol. 90. Elsevier B.V., Apr. 01, 2023. doi: 10.1016/j.ebiom.2023.104512.
31. B. Ji *et al.*, "Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models," *J Biomed Inform*, vol. 104, Apr. 2020, doi: 10.1016/j.jbi.2020.103395.
32. K. Ahmed, S. K. Khurshid, and S. Hina, "CyberEntRel: Joint extraction of cyber entities and relations using deep learning," *Comput Secur*, vol. 136, Jan. 2024, doi: 10.1016/j.cose.2023.103579.
33. X. Li, H. Zhang, and X. H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J Biomed Inform*, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103422.
34. C. Wang *et al.*, "Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree," *J Biomed Inform*, vol. 111, Nov. 2020, doi: 10.1016/j.jbi.2020.103583.
35. C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using BERT in the machine reading comprehension framework," *J Biomed Inform*, vol. 118, Jun. 2021, doi: 10.1016/j.jbi.2021.103799.
36. H. Fabregat, A. Duque, J. Martinez-Romo, and L. Araujo, "Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction," *J Biomed Inform*, vol. 138, Feb. 2023, doi: 10.1016/j.jbi.2022.104279.
37. R. Juez-Hernandez, L. Quijano-Sánchez, F. Liberatore, and J. Gómez, "AGORA: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents," *Appl Soft Comput*, vol. 145, Sep. 2023, doi: 10.1016/j.asoc.2023.110540.
38. N. Jofche, K. Mishev, R. Stojanov, M. Jovanovik, E. Zdravevski, and D. Trajanov, "Named Entity Recognition and Knowledge Extraction from Pharmaceutical Texts using Transfer Learning," in *Procedia Computer Science*, Elsevier B.V., 2022,
39. X. Jiang, Y. Cheng, S. Zhang, J. Wang, and B. Ma, "APIE: An information extraction module designed based on the pipeline method," *Array*, vol. 21, Mar. 2024, doi: 10.1016/j.array.2023.100331.
40. H. Zhang, J. Guo, Y. Wang, Z. Zhang, and H. Zhao, "Judicial nested named entity recognition method with MRC framework," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 118–126, Jun. 2023, doi: 10.1016/j.ijcce.2023.03.002.

41. A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decision Analytics Journal*, vol. 10, Mar. 2024, doi: 10.1016/j.dajour.2024.100426.
42. L. Ding, C. Ouyang, Y. Liu, Z. Tao, Y. Wan, and Z. Gao, "Few-shot Named Entity Recognition via encoder and class intervention," *AI Open*, vol. 5, pp. 39–45, Jan. 2024, doi: 10.1016/j.aiopen.2024.01.005.
43. S. Rizou *et al.*, "Efficient intent classification and entity recognition for university administrative services employing deep learning models," *Intelligent Systems with Applications*, vol. 19, Sep. 2023, doi: 10.1016/j.iswa.2023.200247.
44. B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on Named Entity Recognition — datasets, tools, and methodologies," *Natural Language Processing Journal*, vol. 3, p. 100017, Jun. 2023, doi: 10.1016/j.nlp.2023.100017.
45. Y. Lou, T. Qian, F. Li, and D. Ji, "A Graph Attention Model for Dictionary-Guided Named Entity Recognition," *IEEE Access*, vol. 8, pp. 71584–71592, 2020, doi: 10.1109/ACCESS.2020.2987399.
46. G. Yang and H. Xu, "A Residual BiLSTM Model for Named Entity Recognition," *IEEE Access*, vol. 8, pp. 227710–227718, 2020, doi: 10.1109/ACCESS.2020.3046253.
47. G. Popovski, B. K. Seljak, and T. Eftimov, "A Survey of Named-Entity Recognition Methods for Food Information Extraction," *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 31586–31594, 2020. doi: 10.1109/ACCESS.2020.2973502.
48. T. Saout, F. Lardeux, and F. Saubion, "An Overview of Data Extraction from Invoices," *IEEE Access*, vol. 12, pp. 19872–19886, 2024, doi: 10.1109/ACCESS.2024.3360528.
49. M. Dias, J. Boné, J. C. Ferreira, R. Ribeiro, and R. Maia, "Named entity recognition for sensitive data discovery in portuguese," *Applied Sciences (Switzerland)*, vol. 10, no. 7, Apr. 2020, doi: 10.3390/app10072303.
50. [50] Y. Wang, Y. Sun, Z. Ma, L. Gao, and Y. Xu, "An ERNIE-based joint model for chinese named entity recognition," *Applied Sciences (Switzerland)*, vol. 10, no. 16, Aug. 2020, doi: 10.3390/app10165711.
51. M. A. Alonso, C. Gómez-Rodríguez, and J. Vilares, "On the use of parsing for named entity recognition," *Applied Sciences (Switzerland)*, vol. 11, no. 3. MDPI AG, pp. 1–24, Feb. 01, 2021. doi: 10.3390/app11031090.
52. P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts," *Applied Sciences (Switzerland)*, vol. 11, no. 18. MDPI, Sep. 01, 2021. doi: 10.3390/app11188319.
53. L. Nemes and A. Kiss, "Information extraction and named entity recognition supported social media sentiment analysis during the COVID-19 pandemic," *Applied Sciences (Switzerland)*, vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112211017.
54. Q. Fang, Y. Li, H. Feng, and Y. Ruan, "Chinese Named Entity Recognition Model Based on Multi-Task Learning," *Applied Sciences (Switzerland)*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084770.
55. W. Liu and X. Cui, "Improving Named Entity Recognition for Social Media with Data Augmentation," *Applied Sciences (Switzerland)*, vol. 13, no. 9, May 2023, doi: 10.3390/app13095360.
56. J. Sawicki, M. Ganzha, M. Paprzycki, and Y. Watanobe, "Applying Named Entity Recognition and Graph Networks to Extract Common Interests from Thematic Subfora on Reddit," *Applied Sciences*, vol. 14, no. 5, p. 1696, Feb. 2024, doi: 10.3390/app14051696.
57. L. He, Q. Wang, J. Liu, J. Duan, and H. Wang, "Visual Clue Guidance and Consistency Matching Framework for Multimodal Named Entity Recognition," *Applied Sciences*, vol. 14, no. 6, p. 2333, Mar. 2024, doi: 10.3390/app14062333.
58. F. Yi, B. Jiang, L. Wang, and J. Wu, "Cybersecurity Named Entity Recognition Using Multi-Modal Ensemble Learning," *IEEE Access*, vol. 8, pp. 63214–63224, 2020, doi: 10.1109/ACCESS.2020.2984582.
59. C. M. Tsai, "Stylometric Fake News Detection Based on Natural Language Processing Using Named Entity Recognition: In-Domain and Cross-Domain Analysis," *Electronics (Switzerland)*, vol. 12, no. 17, Sep. 2023, doi: 10.3390/electronics12173676.
60. J. D'Souza, "Agriculture Named Entity Recognition—Towards FAIR, Reusable Scholarly Contributions in Agriculture," *Knowledge*, vol. 4, no. 1, pp. 1–26, Jan. 2024, doi: 10.3390/knowledge4010001.