REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 14 | ISSUE - 3 | DECEMBER - 2024

COMPARATIVE ANALYSIS OF EXISTING ALGORITHMS FOR NER

Mrs. Sneha Desai and Mrs. Monica Parekh

1. ABSTRACT

*Named Entity Recognition (NER) is a foundational task in **Natural Language Processing (NLP)**, aimed at identifying and classifying named entities such as people, organizations, locations, dates, and other entities in unstructured text. Over the years, NER has become integral to numerous real-world applications, including **information extraction**, **question answering**, **machine translation**, and **chatbots**. The effectiveness of NER systems directly influences the performance of downstream NLP tasks, making it a critical area of research.*

*The evolution of NER techniques can be broadly categorized into three generations: **rule-based methods**, **machine learning-based approaches**, and **deep learning-based models**. Rule-based systems rely on hand-crafted rules and dictionaries, which are effective for structured or domain-specific text but struggle with generalization. Traditional machine learning algorithms, such as Conditional Random Fields (CRF) and Support Vector Machines (SVM), improve performance by leveraging statistical learning and feature engineering. However, these methods require extensive manual effort for feature extraction and are limited in their ability to capture complex patterns in text.*

In recent years, the advent of deep learning and transformer-based models has revolutionized NER. Models such as BiLSTM-CRF and pre-trained transformers like BERT and its variants (e.g., RoBERTa, mBERT) have achieved state-of-the-art performance by learning contextual representations of words. These approaches can automatically extract features from data and capture intricate dependencies, making them robust across domains and languages. Despite their success, they come with challenges such as computational overhead, large data requirements, and domain adaptation issues.

This paper presents a comparative analysis of existing algorithms for Named Entity Recognition to highlight their strengths, weaknesses, and applicability across different scenarios. We systematically evaluate traditional machine learning approaches (e.g., CRF, HMM), neural network-based models (e.g., BiLSTM-CRF), and transformer-based architectures (e.g., BERT, RoBERTa) on standard NER datasets.

2. KEYWORDS: *Named Entity Recognition (NER), Transformer-based Models, Domain-Specific Applications, Challenges in NER1Introduction.*

3. INTRODUCTION

Named Entity Recognition (NER) systems face significant challenges when dealing with complex patterns in text. These complexities arise due to the diversity and ambiguity inherent in natural language. Identifying named entities, especially in real-world scenarios, often requires handling a variety of intricate patterns, such as:

1. Ambiguity in Entity Types("Amazon" can refer to a *company* (Amazon.com), a *location* (Amazon rainforest), or a *river* (Amazon River).)
2. Nested Entities("University of California, Berkeley" includes:"University of California" (organization)."Berkeley" (location).)
3. Multi-word Entities("New York City" (location)."The United Nations" (organization).)
4. Entities with Non-standard Spellings or Formats("eBay," "iPhone13," or Elon_Musk.")
5. Language and Domain Variability(Handling languages with complex morphology, such as Finnish or Turkish.)

➢ Handling the Complexity of Language

Ambiguity in Text: Words can have different meanings in different contexts. For instance, "Apple" can refer to a fruit or a company. Algorithms help resolve this ambiguity using contextual clues.

Variety in Entity Types: Entities come in many forms (e.g., names, dates, monetary amounts), and algorithms are required to learn and differentiate between these types effectively.

Diverse Linguistic Structures: Natural language varies in grammar, syntax, and structure. Algorithms can process these variations and identify entities regardless of how they are written.

➢ Scalability

Processing Large Datasets: Algorithms enable automated and efficient processing of vast amounts of text data, which would be infeasible manually.

Real-Time Recognition: In applications like chatbots or information retrieval, algorithms allow NER to be performed in real time, providing immediate responses to users.

➢ Adaptability to Multiple Languages

Language rules and entity formats vary across languages. Algorithms can be trained on multilingual datasets to handle language-specific nuances, making them adaptable to diverse linguistic contexts.

➢ Learning Context and Patterns

Algorithms, especially machine learning models, can identify patterns in data and learn the context in which certain entities occur. For example, statistical models or neural networks can learn that "Dr." is often followed by a person's name.Deep learning models like Transformers (e.g., BERT) are particularly adept at capturing contextual relationships and dependencies within sentences.

➢ Generalization

Algorithms can generalize from training data to new, unseen examples, enabling robust performance across a wide range of real-world scenarios.

➢ Automation and Efficiency

Rule-based systems are limited in scope and require manual updates for every new rule. Algorithms, particularly machine learning-based ones, automate the process by learning from labeled data, reducing the need for extensive manual effort.

➢ Performance and Accuracy

Advanced algorithms like Conditional Random Fields (CRFs), Hidden Markov Models (HMMs), and deep learning models achieve high accuracy in recognizing entities by leveraging statistical and contextual relationships in data.Neural networks, such as recurrent neural networks (RNNs) and transformer models, have significantly improved the accuracy and reliability of NER systems.

➢ **Customizability**

Algorithms can be fine-tuned or retrained for domain-specific applications. For instance, a biomedical NER system might focus on recognizing drug names and diseases, which general-purpose rule-based systems might struggle to handle.

4.OBJECTIVES OF ALGORITHMS FOR NAMED ENTITY RECOGNITION (NER) IN NLP

4.1 Identify and Classify Named Entities:

The primary goal is to detect specific types of entities such as names of people, locations, organizations, dates, quantities, and more.Classification ensures that each detected entity is assigned to the correct category.

4.2 Handle Ambiguity and Context:

Algorithms must disambiguate words based on context. For instance, "Amazon" could refer to a company or a river depending on the surrounding text.

4.3 Support Multiple Languages and Domains:

They aim to generalize across various languages and specialized domains, such as biomedical texts, financial documents, or social media.

4.4 Enable Scalability:

Algorithms should process large datasets efficiently, whether for training on annotated corpora or inference in real-world applications.

4.5 Real-Time and Adaptive Processing:

For applications like chatbots, algorithms need to perform NER in real-time and adapt to dynamically changing contexts or vocabularies.

4.6 Minimize Errors:

Reduce false positives (incorrectly tagging non-entities) and false negatives (failing to recognize valid entities).

5.EFFICIENCY OF ALGORITHMS FOR NER IN NLP

5.1 Computational Complexity: Efficient algorithms balance accuracy with computational resources. For instance:

5.2 Rule-Based Systems: Simple but computationally light, suitable for constrained scenarios.

5.3 Machine Learning Models: Require training but are efficient at runtime.

5.4 Deep Learning Models: Computationally intensive during training, but efficient for inference when optimized.

5.5 Memory Usage: Algorithms like BERT-based models require significant memory for embeddings and model parameters. Optimizations, such as model pruning or quantization, improve efficiency.

5.6 Data Utilization: Algorithms need to make effective use of labeled data to achieve high performance while reducing dependency on extensive annotations through techniques like semi-supervised learning.

5.7 Scalability to Large Datasets: Scalable algorithms process millions of documents efficiently, leveraging parallel processing or distributed computing (e.g., using frameworks like TensorFlow or PyTorch).

5.8 Adaptability to Noise:Efficient algorithms are robust against noisy data, such as spelling errors or informal text in social media.

5.9 Inference Speed:Real-time applications, such as search engines or voice assistants, require fast inference times. Efficient algorithms use optimizations like caching, batch processing, or approximate techniques.

5.10 Energy and Resource Efficiency:Energy-efficient models are critical for deploying NER systems on resource-constrained devices or in environmentally sustainable computing environments.

6. PREVIOUS METHODS FOR NAMED ENTITY RECOGNITION (NER)

6.1. Rule-Based Methods: Early NER systems relied on handcrafted rules and patterns using regular expressions, dictionaries, and grammar rules.Rule-based systems for detecting dates or person names based on fixed patterns (e.g., "Dr." followed by a name).

Advantages: Highly interpretable and easy to implement for specific tasks.

Disadvantages: Lack of scalability, high maintenance, and inability to adapt to diverse contexts or unseen data.

6.2. Statistical Models: Hidden Markov Models (HMMs), Maximum Entropy Models, and Conditional Random Fields (CRFs) became popular in the early 2000s.CRFs, for instance, model sequential data and consider context effectively, making them robust for tasks involving text.

Advantages: Capture dependencies in text, better performance than rule-based systems.

Disadvantages: Require labeled training data, less effective on noisy or out-of-domain data.

6.3.Machine Learning-Based Methods: Supervised learning methods using algorithms like Support Vector Machines (SVMs) and decision trees allowed the automation of entity recognition.

Feature engineering was a critical component, relying on token-level, lexical, and syntactic features.

Advantages: Improved accuracy with fewer manual rules.

Disadvantages: Heavy reliance on feature engineering, struggled with complex relationships in text.

6.4.Deep Learning Approaches: Recurrent Neural Networks (RNNs), especially Bidirectional LSTMs (BiLSTMs), revolutionized NER by learning sequential patterns in text.

BiLSTM-CRF Models combined the strengths of deep learning (contextual understanding) with statistical techniques (structured prediction).Contextual Embeddings like ELMo and BERT improved NER by capturing dynamic word meanings based on context.

Advantages: Handle large datasets, learn complex relationships, and adapt to multiple domains.

Disadvantages: High computational cost, require large amounts of labeled data, and are less interpretable.

7. RESEARCH GAPS IN COMPARATIVE ALGORITHMS FOR NER

Limited Domain Adaptability: Many models are trained on general-purpose datasets (e.g., news articles) and struggle with domain-specific texts such as medical, legal, or financial documents.

Handling Low-Resource Languages: Most state-of-the-art methods focus on high-resource languages like English, neglecting low-resource languages with limited annotated data.

Entity Ambiguity and Context Sensitivity: Current models sometimes fail to distinguish ambiguous entities (e.g., "Amazon" as a river or a company).

Efficiency and Scalability: While deep learning models achieve high accuracy, they often require substantial computational resources, limiting their applicability in real-time or resource-constrained environments.

Noisy and Informal Text: Social media, user-generated content, and other informal text sources introduce challenges like slang, misspellings, and incomplete sentences.

Limited Explainability: Deep learning-based NER models, despite their performance, are often "black boxes," making their predictions less interpretable.

Lack of Multimodal Integration: Most NER algorithms work solely on text data, ignoring other modalities like images or metadata (e.g., a tweet's attached image or location).

8. FUTURE DIRECTIONS IN NAMED ENTITY RECOGNITION (NER):

In the past years many start of the arts approaches were introduced those face many issues, challenges and resolved for the purpose of NER systems. For NER modals highly depend into features selection therefore new clustering techniques are used to address noisy or sparseness data problems.

Supervised machine learning algorithms need a large amount of dataset for training and testing of the models but it is a difficult challenge to collect large number of datasets and annotated from lower resource of languages for instance Pakistan local languages. Unsupervised and Semi-supervised learning algorithms need few amounts of data set which is annotated. Hence, future researchers are focusing to explore semantic related information about words and finding the semantic structure between words for named Entity Recognition systems.

In past studies had extracted only few focused such as name, time location etc. Therefore, new researches are focusing of more fine entities which can help in many information retrieval applications.

In NER systems linguistic features collection is costly, time consuming and memory space issues. So future research can work on statistical approaches apply in NER models for obtaining better results. They are also work to remove the ambiguity in the datasets in NER systems.

Rule-based models are required specific languages and costly, they cannot easily transform into new languages. In other side, machine learning model is not easily portable from one system to another. So that new techniques are combined rule-based and learning based approaches to make for Hybrid approaches which give high quality result and less costly.

9. APPLICATIONS AND ITS CHALLENGES:

The application of Named Entity Recognition (NER) research spans numerous critical domains, each demanding tailored solutions to address distinct challenges and extract domain-specific knowledge effectively. In the realm of healthcare and biomedical research, NER is instrumental in unlocking insights from a burgeoning volume of medical data, enabling advancements in patient care, drug discovery, and disease monitoring. In the cybersecurity domain, NER aids in swiftly identifying and classifying cybersecurity entities, bolstering threat detection and incident response capabilities in the face of evolving cyber threats. Furthermore, NER empowers environmental scientists to decipher climate parameters, species data, and ecological trends, facilitating a deeper comprehension of environmental changes and conservation efforts. In the legal arena, NER streamlines the complex task of identifying legal entities, case references, and legal codes within vast legal documents, enhancing document analysis and compliance monitoring. Meanwhile, NER research in the energy sector supports informed energy policy development and sustainable resource management. During humanitarian crises, NER plays a pivotal role in swiftly locating and coordinating relief efforts for affected populations. In space exploration, NER catalogues celestial discoveries and space mission details, while in the AI and technology sector, it identifies emerging technologies and innovations, informing tech trend analysis and investment decisions. NER in education aids in research, student enrollment, and content recommendation, while in public health, it contributes to timely disease outbreak monitoring and resource allocation during health crises. Across these domains, NER research stands as a crucial and ever-evolving field, continuously adapting to diverse challenges and domains to deliver specialized entity recognition systems that enrich research, industry, and society as a whole.

9.1 Finance Applications:

Named Entity Recognition (NER) has several applications in the finance domain, mainly in the extraction of crucial information from various documents Applications of NER[4] in finance involve extracting important details from documents. For instance, it helps sort transactions by recognizing vendor names, product descriptions, and amounts. By linking a found entities to databases, NER provides extra info, like connecting a company's name to its financial data. It even helps analyze feelings by spotting entities in financial texts, making it possible to gauge opinions about companies, stocks, or events. Moreover, it's used to identify risk-related entities, aiding in evaluating potential financial risks. NER supports compliance with rules by pinpointing entities that matter for regulations. Also, it's handy in managing portfolios by recognizing stock symbols, company names, and financial markers.

- **Challenges:**

However, there are challenges when extracting named entity data from financial documents like PDFs, invoices, and bills. These documents come in different styles, making it hard to consistently find entities. Plus, some documents mix structured info (like tables) with unstructured text, needing special methods. Errors like typos or inconsistent naming in financial documents can affect accurate entity detection. Poor scans could lead to errors, and similar terms may mean different things based on context, needing smart NER models. Financial documents often have multiple entities together (like company names and money amounts), making it tricky to tell them apart. Also, there's not enough labeled data specific to finance, which makes building good models tough. Using specialized words and linking entities to external sources can be tough too, as can ensuring sensitive info is extracted in compliance with rules. Overcoming these issues needs smart NER models, special data, and methods that understand finance's complexities.

9.2 Biomedical Applications:

In the field of Biomedical, Named Entity Recognition (NER) has various uses. It helps in identifying crucial information in medical texts. For instance, it can locate and categorize medical terms like disease names, drug names, and medical procedures. NER also supports medical research by recognizing genes, proteins, and molecular structures mentioned in scientific literature. Moreover, it aids in tracking patient data, linking medical terms with patient records, and helps in managing medical databases.

- **Challenges:**

There are challenges in NER for Biomedical texts. One challenge is the vast number of medical terms and variations, which can make accurate recognition tricky. Also, medical texts often have complex sentence structures and abbreviations that can confuse NER systems. Ambiguities between common words and medical terms pose another issue. Additionally, there's a lack of large annotated datasets specific to Biomedical NER, making training accurate models challenging. Overcoming these hurdles requires advanced algorithms, specialized training data, and methods that can navigate the intricacies of medical language.

9.3 Other applications and challenges

Named Entity Recognition (NER) finds applications across various fields beyond finance and biomedical domains. In legal texts, NER helps identify legal terminologies, case references, and entities like names of laws and regulations. However, challenges arise due to the diverse legal language and context-specific entity meanings. In the news domain, NER aids in extracting people's names, locations, and organization names, supporting news categorization and sentiment analysis. The ambiguity of entity references and rapidly evolving entities pose challenges. Similarly, in the e-commerce sector, NER assists in extracting product names, brands, and specifications. Yet, the vast variety of product names and frequent changes in product listings create difficulties. In social media, NER is used for sentiment analysis, topic identification, and user profiling. However, the informal language, abbreviations, and context-dependent entity mentions make accurate recognition challenging. While Named Entity Recognition(NER) systems excel for English, numerous challenges persist for other Indian and Asian languages. The issues such as the absence of capitalization norms, linguist morphological complexities, the ambiguity between common and proper nouns, and the overlap of terms like person and location names. Overcoming these challenges requires domain-specific [12] training data, context-aware models, and adaptable algorithms.

10. CONCLUSION

In the realm of Natural Language Processing, Named Entity Recognition (NER) stands as a monumental pillar, bridging the gap between unstructured textual data and structured knowledge organization. This comprehensive survey of AI techniques for NER underpins the tremendous strides made in the domain, while also acknowledging the multifaceted challenges it continues to confront. From the rudimentary rule-based methods to the transformative prowess of transformer architectures, the NER landscape has witnessed a profound metamorphosis. The dawn of domain-specific NER models accentuates the adaptability of these techniques, demonstrating the finesse with which models like ViBERTgrid and BioBERT cater to the nuanced requirements of specialized domains like finance and medicine. Yet, the continuous evolution of deep learning paradigms, like E-NER and the utilization of large language models, indicates that the NER journey is still unfolding, with newer horizons awaiting exploration.

The synergy between Optical Character Recognition (OCR) and NER is an exciting testament to the interdisciplinary collaboration within AI, showcasing the potential of integrating diverse technologies for more robust information extraction. Such amalgamations signal the future direction of NER — a direction characterized by holistic, multimodal information extraction that spans not only text but also visual and auditory data.

Despite these advancements, the challenges delineated in the paper underscore the areas awaiting deeper inquiry. The intrinsic hurdles posed by different industries, languages, and the ever-evolving nature of information, emphasize the dynamic terrain of NER. Adapting to low-resource languages and domains, while enhancing the model's ability to handle ambiguity and connect recognized entities to expansive databases, forms the crux of the road ahead. Moreover, as NER cements its role in pivotal applications, from parsing intricate financial documents to deciphering the complexities of medical texts, its societal impact becomes even more pronounced. It becomes imperative to ensure the transparency, reliability, and ethical considerations of these models, considering their profound implications in decision-making processes across sectors.

In summation, while this research offers a panoramic view of the current NER landscape, it also serves as a clarion call to the research community. The realm of Named Entity Recognition is rife with opportunities, challenges, and responsibilities. With the accelerating pace of innovation and the ever-increasing significance of language understanding in the digital age, NER stands not just as a technical endeavor but as a cornerstone in shaping the future of information extraction and knowledge representation. The road ahead is long, promising, and teeming with potential — a journey that beckons the collaborative efforts of researchers, practitioners, and industry stalwarts.

11. REFERENCES:

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.Bert: Pre-training of deep bidirectional transformers for language understanding.*arXiv preprint arXiv:1810.04805*, 2018.
2. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li.A survey on deep learning for named entity recognition.*IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.
3. Girish Keshav Palshikar.Techniques for named entity recognition: a survey.In *Bioinformatics: Concepts, Methodologies, Tools, and Applications*, pages 400–426. IGI Global, 2013.
4. Jason PC Chiu and Eric Nichols.Named entity recognition with bidirectional lstm-cnns.*Transactions of the association for computational linguistics*, 4:357–370, 2016.
5. Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo.Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents.In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 548–563. Springer, 2021.
6. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.Biobert: a pre-trained biomedical language representation model for biomedical text mining.*Bioinformatics*, 36(4):1234–1240, 2020.
7. Dinda Ayu Permatasari and Devira Anggi Maharani.Combination of natural language understanding and reinforcement learning for booking bot.*Journal of Electrical, Electronic, Information, and Communication Technology*, 3(1):12–17, 2021.
8. Yawen Yang, Xuming Hu, Fukun Ma, Aiwei Liu, Lijie Wen, S Yu Philip, et al.Gaussian prior reinforcement learning for nested named entity recognition.In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
9. Liangping Ding, Tian-Yuan Huang, Huan Liu, Yufei Wang, and Zhixiong Zhang.Distantly supervised named entity recognition with category-oriented confidence calibration.In *International Conference on Asian Digital Libraries*, pages 46–55. Springer, 2022.

10. Zhen Zhang, Mengting Hu, Shiwan Zhao, Minlie Huang, Haotian Wang, Lemao Liu, Zhirui Zhang, Zhe Liu, and Bingzhe Wu.E-ner: Evidential deep learning for trustworthy named entity recognition.*arXiv preprint arXiv:2305.17854*, 2023.
11. Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al.Lamda: Language models for dialog applications.*arXiv preprint arXiv:2201.08239*, 2022.
12. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al.Language models are few-shot learners.*Advances in neural information processing systems*, 33:1877–1901, 2020.
13. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al.Palm: Scaling language modeling with pathways.*arXiv preprint arXiv:2204.02311*, 2022.
14. Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel.Comprehend medical: a named entity recognition and relationship extraction web service.In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE, 2019.
15. Amrita Anandika and Smita Prava Mishra.A study on machine learning approaches for named entity recognition.*2019 International Conference on Applied Machine Learning (ICAML)*, pages 153–159, 2019.
16. Antonia Höfer and Mina Mottahedin.Minanto at semeval-2023 task 2: Fine-tuning xlm-roberta for named entity recognition on english data.In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1127–1130, 2023.
17. Roman Dušek, Aleksander Wawer, Christopher Galias, and Lidia Wojciechowska.Improving domain-specific retrieval by nli fine-tuning.*arXiv preprint arXiv:2308.03103*, 2023.
18. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever.Improving language understanding with unsupervised learning.2018.
19. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer.Neural architectures for named entity recognition.*arXiv preprint arXiv:1603.01360*, 2016.
20. Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih.Dissecting contextual word embeddings: Architecture and representation.*arXiv preprint arXiv:1808.08949*, 2018.
21. Chris Alberti, Kenton Lee, and Michael Collins.A bert baseline for the natural questions.*arXiv preprint arXiv:1901.08634*, 2019.
22. Jimmy Lin, Rodrigo Nogueira, and Andrew Yates.*Pretrained transformers for text ranking: Bert and beyond*.Springer Nature, 2022.
23. Anne Stockem Novo and Fatih Gedikli.Explaining bert model decisions for near-duplicate news article detection based on named entity recognition.In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 278–281. IEEE, 2023.
24. Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan.Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.*arXiv preprint arXiv:2305.14450*, 2023.
25. Dipali Baviskar, Swati Ahirrao, and Ketan Kotecha.Multi-layout invoice document dataset (midd): a dataset for named entity recognition.*Data*, 6(7):78, 2021.
26. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.Roberta: A robustly optimized bert pretraining approach.*arXiv preprint arXiv:1907.11692*, 2019.
27. Dhananjay Ashok and Zachary C Lipton.Promptner: Prompting for named entity recognition.*arXiv preprint arXiv:2305.15444*, 2023.
28. Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu.Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search.*arXiv preprint arXiv:2305.12217*, 2023.
29. Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang.Distantly supervised ner with partial annotation learning and reinforcement learning.In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, 2018.

30. Alexander Fritzler, Varvara Logacheva, and Maksim Kretov.Few-shot classification in named entity recognition task.In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000, 2019.
31. Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang.Pixiu: A large language model, instruction data and evaluation benchmark for finance.*arXiv preprint arXiv:2306.05443*, 2023.
32. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.Llama: Open and efficient foundation language models.*arXiv preprint arXiv:2302.13971*, 2023.
33. Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei.Document ai: Benchmarks, models and applications.*arXiv preprint arXiv:2111.08609*, 2021.
34. Anand Kumar, Pardeep Singh, and Kusum Lata.Comparative study of different optical character recognition models on handwritten and printed medical reports.In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 581–586. IEEE, 2023.