An Empirical Study of Named Entity Recognition for Electronic Health Records

Nichapha Manoonwong, Anca Muresan, Anudeep Reddy Raavi, and Xingquan Zhu
Department of Electric Engineering and Computer Science
Florida Atlantic University
Boca Raton, FL-33431, USA
{nmanoonwong2020, amuresan2023, ranudeep2023, xzhu3}@fau.edu

2024 IEEE International Conference on Knowledge Graph (ICKG) | 979-8-3315-0882-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICKG63256.2024.00036

***Abstract*—Named Entity Recognition (NER), which locates and recognizes phrases (*i.e.,* entities) carrying specific meaning, such as locations, organization names, *etc.*, is an important information extraction step for domains such as the processing of biomedical or Electronic Health Records (EHRs). To date, many methods exist for NER, and neural language models have emerged as the most sought-after tool due to their superb performance compared to other alternatives, such as rule-based approaches. Nevertheless, the rich set of tools and algorithms also raises concerns for both researchers and practitioners: how algorithms vary in their performance and what the general practices are in selecting proper NER algorithms for EHR processing. In this paper, we conduct an empirical study to understand the performance of different types of neural language models for EHR named entity recognition. Our main goal is to infer the performance of each model type with respect to the sample volumes and distributions with a special emphasis on context-dependent and low-frequency entities that pose a significant challenge, even for state-of-the-art models. Five types of models, LSTM, BiLSTM, Basic BiLSTM-CRF, Enhanced BiLSTM-CRF, and BERT, are studied in our experiments by using the N2C2 dataset (unstructured notes from the research patient data registry) as the test-bed. We vary the sample volumes and distributions and comparatively study the model performance. Our study draws important findings for researchers to decide the most suitable NER tools for EHRs.**

***Index Terms*—Named entity recognition, neural language models, electronic health records, information extraction**

I. Introduction

Named Entity Recognition (NER) [1], [2] is one of the core tasks in Natural Language Processing (NLP) that involves extracting and classifying distinct words or phrases from text, called "entities" [3]. These entities can represent predefined categories such as locations, organizations, dates, or other categories referring to a particular domain [4]. In the medical field, NER is essential for extracting meaningful information from vast unstructured datasets like clinical EHRs [5]. For example, NER can be used to identify drug names, diseases, symptoms, and diagnoses [6], which helps summarize key details in patient reports. This information can streamline various healthcare operations, including medication dispensing systems and patient tracking for follow-up appointments, and enhance the overall efficiency of the healthcare system through the revolutionary capability of clinical NLP [7].

There are several approaches to NER. Traditional rule-based methods [8] have been outperformed by more advanced supervised learning algorithms. In supervised NER, common methods include Hidden Markov Models (HMM) [9], Conditional Random Fields (CRF), and neural network-based models like Bi-directional Long Short-Term Memory (BiLSTM) [10]. These models are particularly effective at capturing sequential dependencies in text, which is crucial for handling complex medical contexts. Another widely used approach involves pre-trained neural language models, such as Bidirectional Encoder Representations from Transformers (BERT) [11], [12]. BERT models, when specifically fine-tuned for biomedical text, have shown strong performance in task like Medical Named Entity Recognition (MedNER) [13].

Despite the progress in NER methodologies, several challenges persist. One of the major issues in learning-based approaches is the quality and distribution of the training data [14]. Imbalanced datasets [15], where certain entity types are underrepresented, can introduce biases in the model's predictions. Additionally, noisy or inconsistent labeling in the data can degrade model performance [16]. In MedNER, the complexity of certain categories such as drugs, procedures, and adverse events makes it more challenging for models to accurately learn and identify these entities.

In this study, we explore and develop NER for EHRs, focusing on drugs and their related entities. We use the 2018 N2C2 shared task dataset [17] as our benchmark. The N2C2 dataset has been frequently used as a baseline in many studies due to its rich annotations and comprehensive coverage of clinical concepts. Studies such as [18] and [19] have relied on it to evaluate and improve models for tasks like medication and ADE extraction, highlighting its importance as a reliable resource for benchmarking NER models in clinical settings. An example of named entities from medical notes is shown in Fig. 1. However, while this dataset provides a strong foundation, relying solely on it may limit the generalizability of our findings to other datasets or real-world clinical settings.

We compare the performance of five neural models: LSTM (baseline), BiLSTM, Basic BiLSTM-CRF, Enhanced BiLSTM-CRF, and BERT, across different sample volumes and distributions. We aim to evaluate their strengths and weaknesses in recognizing medical entities, offering a comparative

979-8-3315-0882-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICKG63256.2024.00036

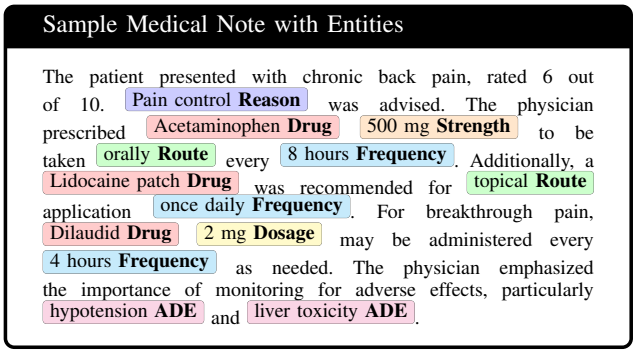

Fig. 1: Example of named entities from medical notes. Text highlighted denotes identified named entities. Entities are color-coded with each color denoting entities of the same type. For each entity, the texts in normal format are the actual named entity, and the bold-faced text denotes the type of the entity.

baseline for model performance. This analysis lays the groundwork for future research involving more diverse datasets and helps guide researchers and practitioners in selecting effective NER models for EHR processing.

II. Related Work

A. Generic Named Entity Recognition

In generic named entity recognition, the evolution of NER techniques spans from rule-based methods to advanced machine learning and deep learning models [20]. Initially, Rule-Based Approaches were prevalent, utilizing extensive dictionaries and simple heuristics but struggled with the complexity and variability of terms, especially in non-standardized data like social media or informal texts. Transitioning from these, Machine Learning-Based Approaches such as HMM and CRF began to show promise by capturing more contextual information and handling sequential data effectively [21].

The advent of Deep Learning-Based Approaches marked a significant shift towards models like BiLSTM, which excel at understanding long-range dependencies within text, crucial for accurate entity recognition across diverse datasets, including large-scale internet-sourced data. This subsection naturally leads to discussing Pre-Trained Language Models, such as BERT integrated with BiLSTM and CRF, which leverages deep contextualized embeddings to further enhance NER performance, showing substantial improvements in robust datasets and achieving high accuracy in complex entity scenarios [22].

B. Domain Specific Named Entity Recognition

Domain-specific NER methods leverage specialized knowledge to enhance the underlying NER modules, with Rule-Based Approaches serving as an early foundational strategy. Though simple, these methods are powerful when enriched with domain-specific dictionaries and rule sets tailored to specialized fields such as biomedicine. By focusing on precise terminologies, such as protein entities, and augmenting dictionaries with POS tagging [23], these systems improve recognition accuracy within highly technical texts, particularly in specialized domains like biomedical literature.

Machine Learning-Based Approaches, such as CRF [24], have been employed to refine the entity identification process by considering both local features and the broader contextual relationships within the text. This capability makes CRF models particularly effective in domains where understanding context is essential, such as in biomedical datasets, where the relationships between terms and entities are crucial to accurate recognition [25].

Deep Learning-Based Approaches, particularly BiLSTM, have demonstrated significant potential in analyzing complex datasets by leveraging their ability to capture sequential dependencies. When adapted for domain-specific tasks, such as medical NER, these models benefit from the integration of domain-specific embeddings, leading to notable improvements in performance metrics [26]. One notable example is a BiLSTM-CRF hybrid model, enhanced with Word2Vec embeddings specifically trained on medical data, developed for a medical virtual assistant system [27]. This advanced setup effectively differentiates between medical and non-medical terminology in real-time consultations, showcasing the model's ability to manage specialized vocabulary in dynamic environments.

Another significant advancement in medical NER involves enhancing BiLSTM-CRF models with character-level word vectors and pre-trained embeddings [28]. By leveraging these features, the model can extract detailed patient data, including symptoms, test results, and treatments, demonstrating an ability to manage complex medical information effectively. This refinement in the model's architecture allows for a more granular understanding of medical texts, enhancing its precision in identifying entities across various medical datasets. Moreover, fine-tuning approaches have proven effective, where models trained from scratch with CRF layers can perform comparably to pre-trained models, such as BERT, highlighting the flexibility and adaptability of these frameworks for domain-specific tasks [29].

Finally, Pre-Trained Language Model-Based approaches have revolutionized domain-specific NER tasks. [30] Models such as BioBERT, pre-trained on vast amounts of biomedical literature, exhibit a deep understanding of domain-specific nuances, allowing them to excel in tasks like medical NER, where recognizing complex interrelationships between medical terms is crucial. These models consistently outperform traditional machine learning and deep learning methods, demonstrating their adaptability and superior performance in specialized fields.

III. Methodology

A. Overall Framework

The overall framework diagram, illustrated in Figure 2, presents our approach used to assess and compare the effectiveness of five distinct NER models. Starting with the initial dataset and moving through stages of preprocessing and sampling to prepare the training data. Then, the data feeds into several model architectures—LSTM, BiLSTM, two

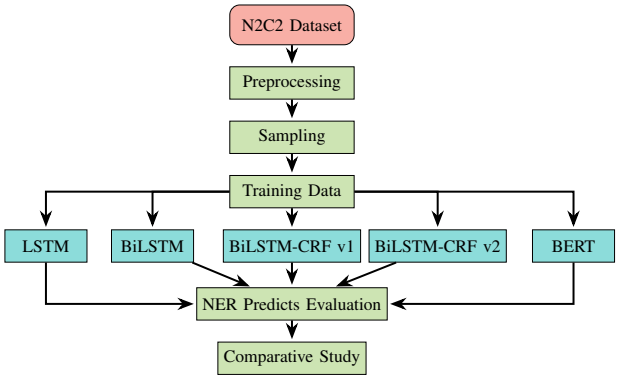

Fig. 2: Overall framework diagram for comparative study between different NER models

versions of BiLSTM-CRF, and BERT. These models are assessed through an NER performance evaluation, followed by a comparative analysis to identify the most effective model for classifying medical terms.

B. Dataset

The N2C2 2018 dataset consists of clinical notes and annotations for drug-related information. This dataset was specifically designed for the N2C2 shared task on medication and ADE extraction, making it a widely used benchmark in medical NER research. The dataset contains 303 training files and 202 test files, annotated with nine entity labels: Drug, Strength, Form, Frequency, Route, Dosage, Reason, Duration, and ADE (Adverse Drug Event), as shown in Table I.

Table II shows the token frequencies for each label, highlighting significant class imbalances in the dataset. There is a high volume of tokens for the Drug and Frequency entities, whereas ADE and Duration are much less frequent.

TABLE I: Named Entity Master

Entity	Description	Example Data
Drug	The product name of the medicine or a chemical substance name.	Aspirin, Tylenol Elixir, oxazepam, Lorazepam, metoprolol succinate
Frequency	The rate at which a drug should be taken in a specific period of time.	DAILY, every eight (8) hours, Q24H, every day
Strength	The amount of chemical of a drug in a given Dosage.	15 mg, 2mg/mL, 200 mg, 10 units/ml, 4200 units.
Form	The Form in which a drug is used.	Pill, Solution, tablet(s), amps, Capsule.
Dosage	The volume of drugs the patient should take.	One (1), several units, 1, 10mL, 100 units/ml.
Reason	The reason for the administration of drugs.	CAD, volume overload, agitation, wheeze, aspiration/pneumonia
Route	The way in which a drug is given into a body or the location of absorption of a drug into the body.	Gtt, Injection, by mouth, PO, Inhalation.
ADE	The development of adverse or unfavorable effects due to drug intake.	Cardiomyopathy, diarrhea, intoxication, morbilliform drug rash
Duration	The length of period of time a drug should be taken.	for at least 1 year, 14 day, for 10 days, two week, 8 day.

TABLE II: Token statistics of the n2c2 dataset

Entity	Train Tokens	Test Tokens	Total Tokens	Percentage
Drug	19,241	12,507	31,748	24.48%
Frequency	15,092	9,616	24,708	19.05%
Strength	12,096	7,629	19,725	15.21%
Form	9,070	6,044	15,114	11.65%
Dosage	7,445	4,783	12,228	9.43%
Reason	6,786	4,556	11,342	8.75%
Route	5,775	3,752	9,527	7.35%
ADE	1,704	1,081	2,785	2.15%
Duration	1,545	973	2,518	1.94%
Total	**78,754**	**50,941**	**129,695**	**100.00%**

C. Experiment Settings

To prepare the data for training, we initially standardized all text in lowercase to ensure uniformity across the dataset. We then tokenized the text using the Sci-Spacy tokenizer, specifically optimized for biomedical text, which is crucial for capturing the domain-specific nuances of the medical language. During the preprocessing step, each tokenized word was transformed into a dense vector using pre-trained word embeddings. These embeddings are vital because they provide a rich representation of the semantic meaning of each word, which is essential for accurate recognition of entities in the medical domain. We inputted these embeddings into various NER models, including LSTM, BiLSTM, Basic BiLSTM-CRF, Enhanced BiLSTM-CRF, and BERT, each fine-tuned to classify entities effectively from their representations.

In our study, we explored the effectiveness of these five different models using training subsets of the dataset containing 10, 25, 100, 200, and 303 data files. This evaluation served two purposes: first, to compare the performance across the five models, and second, to assess how varying amounts of training data influence the performance of each model.

D. LSTM (model 1)

The long short-term memory LSTM model Figure 3 is designed to process sequence data for NER tasks. The model architecture begins with an input layer that feeds text data into an embedding layer, transforming words into dense vector embeddings that encapsulate semantic information. This embedding output is then processed by an LSTM layer designed to capture temporal dependencies in the data. Following the LSTM layer, a dropout of 0.5 is applied to mitigate overfitting by randomly omitting features during training. The sequence output from the dropout layer is passed through a time-distributed dense layer with softmax activation, assigning probabilities to each class for each time step and facilitating the final entity classification.

E. BiLSTM (model 2)

The BiLSTM model Figure 4 uses a deep learning framework to enhance the processing of sequential text data for the recognition of named entities. The model begins with an input layer that feeds data into an embedding layer, converting tokens into dense vector representations. Following embedding, the data are processed through two consecutive BiLSTM

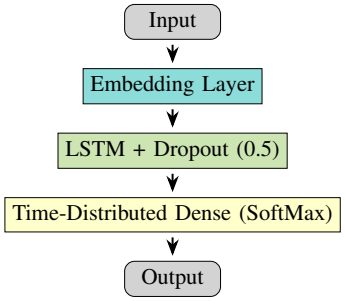

Fig. 3: The architecture of the LSTM (model 1)

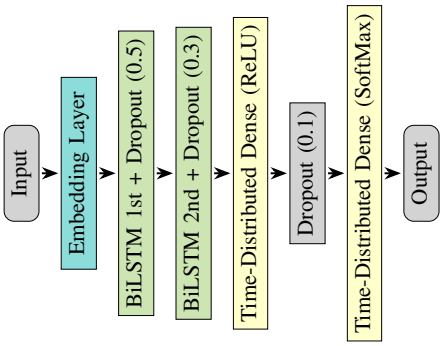

Fig. 4: The architecture of the BiLSTM (model 2)

layers. Each BiLSTM layer is followed by a dropout layer, first with a dropout rate of 0.5 and then 0.3. The output of the BiLSTM layers is then passed through time-distributed dense layers. The first dense layer applies a ReLU activation function and L2 regularization to enhance generalization, followed by another dropout layer with a rate of 0.1. The final layer is another time-distributed dense layer that employs a softmax activation to classify each token into entity categories.

F. BiLSTM-CRFs (Basic model 3, Enhanced model 4)

The BiLSTM-CRF models are designed to enhance NER task by integrating BiLSTM layers with a CRF layer. These models leverage the sequence processing capabilities of BiLSTM layers and the advanced sequence labeling provided by the CRF layer, which uses transition and emission scores to model dependencies between consecutive labels in a sequence. The Basic BiLSTM-CRF Model Figure 5 consists of an embedding layer followed by a single BiLSTM layer and a CRF layer. This configuration suits scenarios where simpler model architectures are sufficient to achieve good performance. In contrast, the Enhanced BiLSTM-CRF Model Figure 6 builds on this by including an additional BiLSTM layer and placing dropout layers between them for improved regularization and robustness. This setup is intended for more complex NER tasks that benefit from a deeper learning architecture.

G. BERT (model 5)

The BERT model, based on the Transformer architecture, utilizes encoder layers, as illustrated in Figure 7. The model begins processing with two special tokens, [CLS] at the start

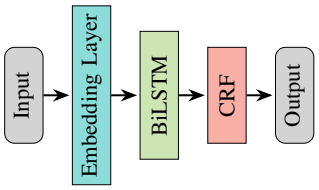

Fig. 5: The architecture of Basic BiLSTM-CRF (model 3)

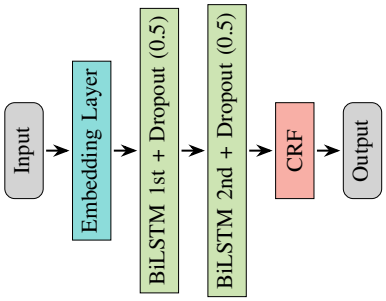

Fig. 6: The architecture of Enhanced BiLSTM-CRF (model 4)

and [SEP] at the end of the input, to handle whole sentences simultaneously through parallel processing. Each encoder layer comprises feedforward networks and multiple attention heads that work together to generate contextualized embeddings from the input text. These embeddings pass sequentially through each encoder layer, with each layer refining the embeddings by capturing different linguistic aspects, such as word meanings and their relationships within the sentence. After progressing through all the encoder layers, the final embeddings can be used by a classifier to accurately label each word in its respective category. This process allows BERT to understand and represent the context of each word within the sentence effectively.

IV. Results

A. Baseline Results

The LSTM (model 1) serves as our baseline method, providing a fundamental point of comparison for more advanced architectures evaluated in this study. This model utilizes a simpler approach to processing textual data, where it transforms input text into semantic embeddings, processes them through an LSTM layer to capture temporal relationships, and then classifies each token based on the learned context.

The performance of this baseline model is presented in the classification report shown in Table III. These metrics serve as initial benchmarks for comparing more sophisticated model architectures (Models 2-5). The results show that ADE has the lowest F1 score of 0.55, likely due to its low number of instances. However, despite Drug being the most frequent entity, it does not achieve the highest score. Instead, Frequency reaches the highest F1 score of 0.99.

Additionally, the performance of the LSTM baseline model across different training file sizes for each entity is illustrated in Figure 8. As expected with a highly imbalanced dataset, the

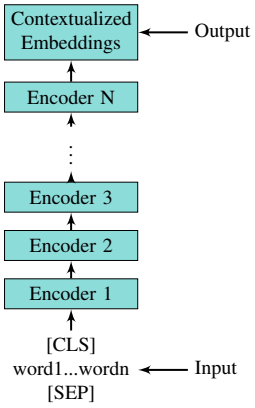

Fig. 7: The process overview of BERT (model 5)

performance on rare entities, such as ADE and Duration, is significantly poor when using smaller subsets of data. While performance improves as more training data is utilized, the low-frequency entities are not predicted at all until training with at least 100 files, and only show slight improvement with additional data. Even with the full 303 training files, the model still struggles to adequately capture these low-frequency entities.

B. Proposed Model Results

In this section, we compare the performance of the different models evaluated in the study: LSTM (baseline), BiLSTM, Basic BiLSTM-CRF, Enhanced BiLSTM-CRF, and BERT. For each model, the F1 scores across the nine entity types—Drug, Form, Strength, Frequency, Route, Dosage, Reason, ADE and Duration—are plotted for training subsets containing 10, 25, 100, 200, and 303 files. These charts provide insight into how well each model performs across entity types as the amount of training data increases.

1) BiLSTM (model 2)

Figure 9 shows the performance of the BiLSTM model. There is a significant improvement in the performance of low-instance classes, such as ADE and Duration. ADE starts with an F1 score of 0.02 and jumps to 0.3 as the training dataset increases from 10 to 25 files, continuing to improve with more data. Similarly, Duration starts at 0.01 and jumps to 0.64, showing consistent progress. In contrast, more frequent entities

TABLE III: Classification Report for Baseline Model

Label	Precision	Recall	F1-score
ADE	0.57	0.53	0.55
Dosage	0.93	0.94	0.94
Drug	0.97	0.95	0.96
Duration	0.94	0.79	0.86
Form	0.98	0.97	0.97
Frequency	1.00	0.98	0.99
Reason	0.88	0.73	0.80
Route	0.98	0.96	0.97
Strength	0.80	0.98	0.88
accuracy			0.93
macro avg	0.89	0.87	0.88
weighted avg	0.94	0.93	0.93

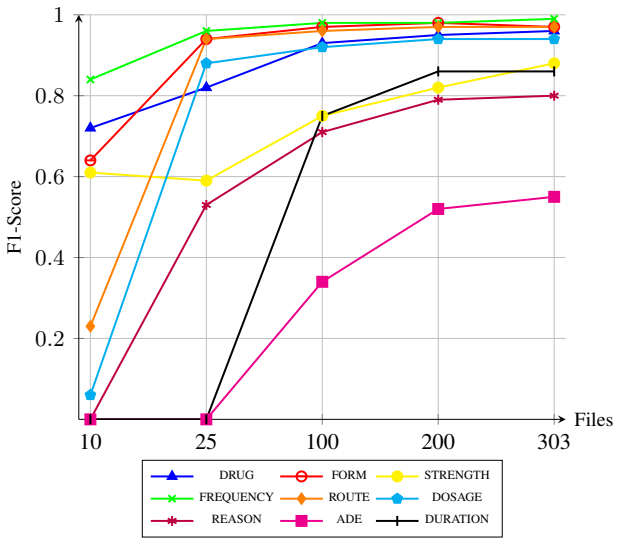

Fig. 8: The F1-Score of the LSTM model (Model 1) with respect to different training set size

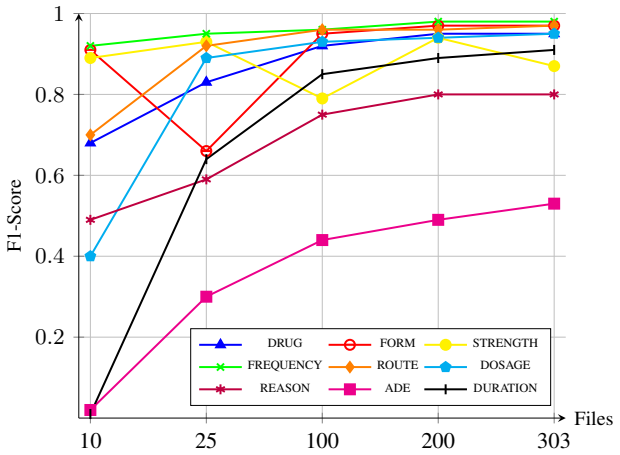

Fig. 9: The F1-Score of the BiLSTM model (Model 2) with respect to different training set size

like Frequency and Strength show only slight improvement across the different dataset sizes.

2) Basic BiLSTM-CRF (model 3)

Figure 10 shows the performance of the model. With the addition of CRF layers, the model is better able to capture the data, as seen in the improved scores for lower-frequency classes, even with as few as 10 files. However, the improvement becomes less significant after training with 100 files, showing only marginal gains with more data.

3) Enhanced BiLSTM-CRF (model 4)

With an additional BiLSTM layer, the performance of the Enhanced BiLSTM-CRF model is shown in Figure 11. The overall trends are similar, with slight improvements in some labels and slight declines in others between 10 and 200 files. These changes do not appear to be related to the frequency of the labels. However, when trained on the full dataset, the

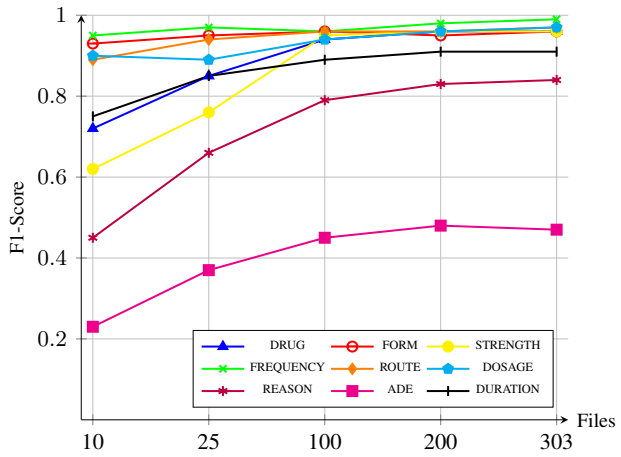

Fig. 10: The F1-Score of the basic BiLSTM-CRF model (Model 3) with respect to different training set size

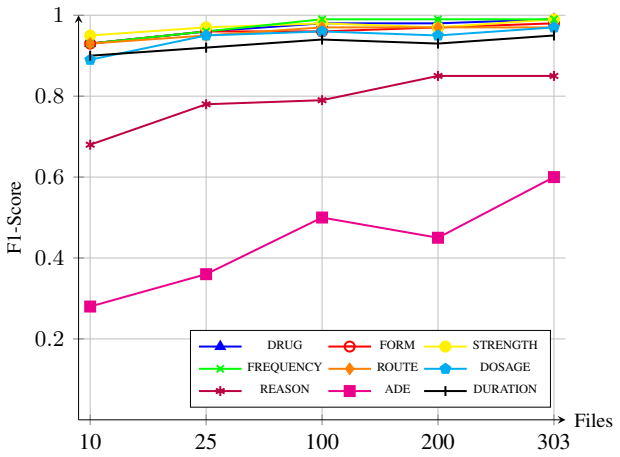

Fig. 12: The F1-Score of the BERT (Model 5) with respect to different training set size

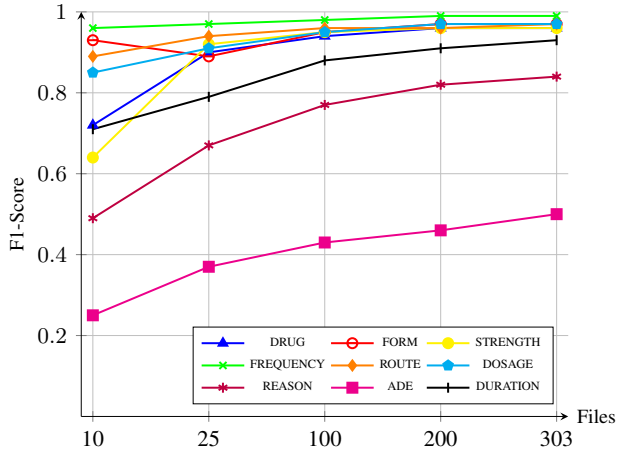

Fig. 11: The F1-Score of the enhanced BiLSTM-CRF model (Model 4) with respect to different training set size

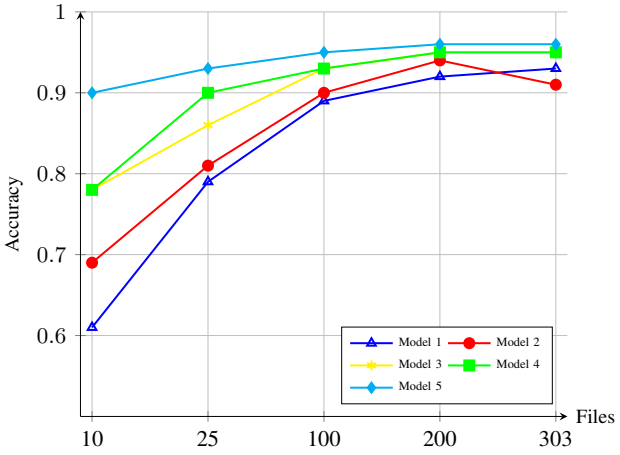

Fig. 13: Accuracy comparison of all models with respect to the size of the training data

model achieves higher scores across all entities compared to model 3, which shares the same structure but has fewer BiLSTM layers.

4) BERT (model 5)

Figure 12 shows the performance of the pre-trained BERT model. Most entities, except ADE and Frequency, achieve very high scores from the beginning, even when trained with only 10 files. There is no significant improvement as the dataset size increases to 303 files. The Frequency entity improves from 0.93 to 0.99, while ADE entity shows more noticeable improvement, starting at 0.28 and reaching 0.6 with the full training data.

C. Model Performance Comparison

Figure 13 shows the overall accuracy for each model across different training dataset sizes. Model 5 consistently achieved the best performance, reaching an accuracy of 0.95 when trained with the full dataset, followed by Model 4 and Model 3, both reaching 0.95 as well. The baseline model (Model 1) initially had the lowest performance but improved as the dataset size increased, ultimately outperforming Model 2 with a final accuracy of 0.93, compared to 0.91 for Model 2.

It is worth noting that both Model 3 and Model 4 were designed to address class imbalance and showed promising results in improving performance. However, their effectiveness in dealing with this issue was not quite similar, suggesting that the additional BiLSTM layer in Model 4 did not result in significant improvements over the single-layer BiLSTM architecture of Model 3 for this particular task.

In our analysis of the imbalanced class, we selected several entities to display in Figure 14, which shows the F1-scores when trained on the entire dataset. The most frequent entity, "Drug", and the less frequent entities, "Reason," "ADE", and, "Duration" were specifically highlighted ("Drug" is the most frequent entity, "Reason" is the 6^{th} most frequent entity, and "ADE" and "Duration" are the top-2 least frequent entities). These low-frequency entities present significant challenges for accurate prediction, as noted in the existing literature.

For the high-frequency entity, "Drug," all models performed exceptionally well, with F1-scores reaching as high as 0.99. Nevertheless, for the low-frequency entities, such as ADE, all models' performance deteriorate significantly. Meanwhile, the performance for low-frequency entities improved slightly compared to the baseline model but still posed challenges.

Interestingly, despite of relatively low frequency, the entity "Reason" and "Duration" both achieved good F1-scores among the minority entities. We believe that this might be attributed to their unique patterns which make entity recognition relatively easier to recognize them. For example, "Duration" entities are frequently associated with a numeric number followed by a date type, such as year, day, week, *etc.*. Meanwhile, "Reason" entities also frequently associate with medical symptoms, making them relatively easier to recognize.

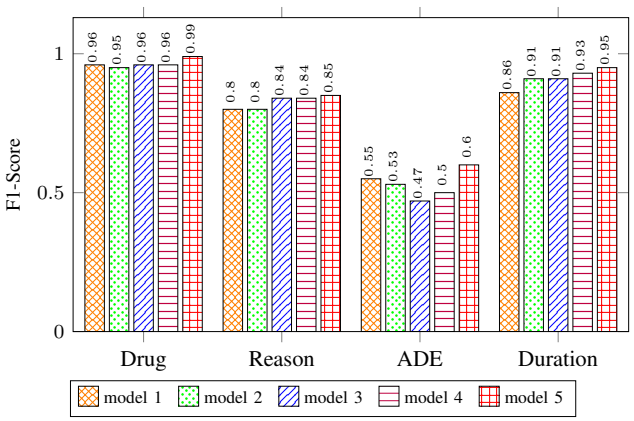

Fig. 14: Comparison of model performance across different categories. "Drug" is the top most frequent entity, "Reason" is the middle frequent entity, "ADE" and "Duration" are the top two least frequent entities, respectively

V. Conclusions

This project aimed to explore and compare the performance of various NER models on EHRs, focusing on the impact of dataset size on model performance. Four models were developed, along with one pre-trained model. Overall, the BiLSTM, BiLSTM-CRFs, and BERT models outperformed the baseline LSTM model, with BERT achieving the highest accuracy. This result is attributed to BERT's extensive pre-training on large corpora, enabling it to capture complex language patterns and domain-specific nuances more effectively.

While BERT showed only marginal improvements with additional task-specific data, the other models exhibited significant performance gains as the amount of training data increased. However, class imbalance remained challenging for all models. High-frequency entities, such as "Drug" consistently yielded better results than low-frequency entities like "ADE". The integration of CRF layers in the BiLSTM models improved the ability to capture contextual relationships, particularly for minority classes, although the difference in performance between the single-layer and multi-layer BiLSTM-CRF models was minimal.

Future research could focus on addressing the class imbalance issue more effectively. Approaches like class weighting and data eesampling could potentially enhance the performance of models on low-frequency entities. Additionally, experimenting with domain-specific models, such as BioBERT, could provide valuable insights into how pre-trained models adapt to the unique challenges of EHR data.

Acknowledgements

This work has been supported in part by the US National Science Foundation (NSF) under Grant Nos. IIS-2236579, IIS-2302786, and IOS-2430224.

References

[1] A. Anandika and S. P. Mishra, "A study on machine learning approaches for named entity recognition," in *2019 International Conference on Applied Machine Learning (ICAML)*, 2019, pp. 153–159.

[2] K. Pakhale, "Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges," *arXiv preprint arXiv:2309.14084*, 2023.

[3] A. Mansouri, L. S. Affendey, and A. Mamat, "Named entity recognition approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339–344, 2008.

[4] R. Sharma, D. Chauhan, and R. Sharma, "Named entity recognition system for the biomedical domain," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2022, pp. 837–840.

[5] L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, "Electronic health records: Implications for drug discovery," *Drug discovery today*, vol. 16, no. 13-14, pp. 594–599, 2011.

[6] P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel, "Comprehend medical: A named entity recognition and relationship extraction web service," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 1844–1851.

[7] M. Bekbolatova, J. Mayer, C. W. Ong, and M. Toma, "Transformative potential of ai in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives," *Healthcare*, vol. 12, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2227-9032/12/2/125

[8] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[9] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *Proceedings of the 40th annual meeting of the association for computational linguistics*, 2002, pp. 473–480.

[10] R. Zhang, P. Zhao, W. Guo, R. Wang, and W. Lu, "Medical named entity recognition based on dilated convolutional neural network," *Cognitive Systems Research*, 2021. [Online]. Available: https://doi.org/10.1016/j.cogr.2021.11.002

[11] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] M. V. Koroteev, "Bert: A review of applications in natural language processing and understanding," 2021. [Online]. Available: https://arxiv.org/abs/2103.11943

[13] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman, and M. S. Haque, "Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts," in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, Torino, Italy, 2023, pp. 158–162.

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[15] S. Nemoto, S. Kitada, and H. Iyatomi, "Majority or minority: Data imbalance learning method for named entity recognition," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2401.11431

[16] S. R. Kundeti, J. Vijayananda, S. Mujjiga, and M. Kalyan, "Clinical named entity recognition: Challenges and opportunities," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1937–1945.

[17] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 3–12, 2020.

[18] Y. Kim and S. M. Meystre, "Ensemble method-based extraction of medication and related information from clinical texts," *Journal of the American Medical Informatics Association : JAMIA*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195830454

[19] F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa, and S. Ananiadou, "Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 39–46, 08 2019. [Online]. Available: https://doi.org/10.1093/jamia/ocz101

[20] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on named entity recognition – datasets, tools, and methodologies," *Natural Language Processing*, 2023. [Online]. Available: https://doi.org/10.1016/j.nlp.2023.100017

[21] Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou, "How to make the most of ne dictionaries in statistical ner," *BMC Bioinformatics*, vol. 9, no. 11, p. S5, 2008, published: 2008/11/19. [Online]. Available: https://doi.org/10.1186/1471-2105-9-S11-S5

[22] W. Li, Y. Du, X. Li, X. Chen, C. Xie, H. Li, and X. Li, "Ud_bbc: Named entity recognition in social network combined bert-bilstm-crf with active learning," *Engineering Applications of Artificial Intelligence*, 2022. [Online]. Available: https://doi.org/10.1016/j.engappai.2022.105460

[23] M. Straka and J. Straková, "Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, J. Hajič and D. Zeman, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 88–99. [Online]. Available: https://aclanthology.org/K17-3009

[24] R. Panchendrarajan and A. Amaresan, "Bidirectional lstm-crf for named entity recognition." 32nd Pacific Asia Conference on Language, Information and Computation, 2018.

[25] A. Thukral, S. Dhiman, and R. Meher, "Knowledge graph enrichment from clinical narratives using nlp, ner, and biomedical ontologies for healthcare applications," *International Journal of Information Technology*, vol. 15, pp. 53–65, 2023. [Online]. Available: https://doi.org/10.1007/s41870-022-01145-y

[26] S. Bruhns, *An empirical study of performance metrics for classifier evaluation in machine learning*. Florida Atlantic University, 2008.

[27] J. Ji, B. Chen, and H. Jiang, "Fully-connected lstm–crf on medical concept extraction," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 1971–1979, 2020. [Online]. Available: https://doi.org/10.1007/s13042-020-01087-6

[28] M. Belousov, N. Milosevic, G. Alfattni, H. Alrdahi, and G. Nenadic, "Gnteam at 2018 n2c2: Feature-augmented bilstm-crf for drug-related entity recognition in hospital discharge summaries," 2019. [Online]. Available: https://arxiv.org/abs/1909.10390

[29] S. Belkadi, L. Han, Y. Wu, and G. Nenadic, "Exploring the value of pre-trained language models for clinical named entity recognition," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 3660–3669.

[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682