SkinNER: Named Entity Recognition (NER) for skin diseases

Rakesh Santha Kumaran
Indiana University
rsantha@iu.edu

Sarabesh Neelamegham Ravindranath
Indiana University
sneelam@iu.edu

Abstract

Named entity recognition (NER) is a natural language processing task that has wide application and serves as a foundation for other tasks such as information retrieval, subject modeling, and question answering. NER is essential in the medical field since it extracts significant sections from clinical notes and reports. These sections are then fed into tasks that come after, such as entity resolution, assertion status detection, relation extraction, and de-identification. There are many NER models for use in the medical domain, but most of them are trained on a corpus containing data about the most common disease.

In this paper, we have worked on SkinNER, a NER model for the identification of all types of skin diseases in a given text including hair and nail-related ones. Tagging these specific diseases which get missed out in existing NER models for diseases, effectively helps the corresponding healthcare professionals to analyze the trends and helps in automatic documentation, and surveillance of reports and records, for further data mining.

1. Introduction

Dermatology is a broad specialty that treats a wide range of skin diseases, from common ones like eczema and acne to rare and complicated illnesses with unusual clinical presentations. Even though a great deal of progress has been achieved in the development of computational models for dermatological applications, there is still a big gap in the analysis and treatment of numerous uncommon skin illnesses.

The models that are now in use frequently depend on datasets that mostly include well-researched and common disorders, which causes them to ignore the wide range of uncommon dermatoses. Consequently, a significant number of skin illnesses remain understudied and underrepresented because the data included in these models do not accurately reflect the broad terrain of dermatological pathology.

This is a serious omission, especially for the many people who suffer from uncommon skin conditions. These illnesses, which frequently have illusive symptoms, slow diagnosis times, and few available treatments, place a significant burden on both patients and healthcare systems. It is estimated that around 7 million persons are affected by rare skin diseases globally at any time. Many rare skin illnesses are still largely unknown in medical literature and computational frameworks, although impacting thousands of people worldwide.

Given these factors, it is critical to close this representation gap in dermatological data and make sure that computer models accurately represent the entire range of skin conditions, particularly those that are uncommon or infrequently reported. We can only hope to meet the unmet needs of the millions of people afflicted by rare skin disorders and advance the science of dermatology toward more inclusive and thorough methods of diagnosis, treatment, and research once we address this deficiency.

To meet the pressing need for more thorough data representation in computational models for dermatology, a Named Entity Recognition (NER) model trained on particular skin-related disorders from medical data corpus such as PubMed records becomes an essential answer. NER is a core task in natural language processing (NLP) that entails locating and classifying named entities (e.g., illnesses, places, and entities of interest) within textual data. A NER model specifically designed to identify and categorize skin-related disorders has great potential to enhance computational frameworks with a more representative and varied dataset in the field of dermatology.

NER entails labeling each entity of interest with its appropriate category and using annotated text data to train a machine learning system. To automatically identify similar entities in unseen text, the model first learns to recognize patterns and linguistic features associated with

these things.

We have created an expert NER system called SkinNER to address this requirement and identify disorders related to the skin. Two separate models make up SkinNER: one is trained from scratch using Spacy NER, while the other is optimized using a BERT architecture that has already been trained. In Section 4, we will go into great detail about these models, including their architecture, training process, and performance assessment of both models. And in Section 6, their performance will be compared and evaluated.

2. Related work

In exploring related works, my paper parallels research efforts that leverage Named Entity Recognition (NER) techniques across various medical domains. A significant point of reference is an example that highlights the use of NER for tagging ocular illnesses and creating a registry specifically for them. This work, cited as [1], investigates a range of model architectures to perform NER tasks on ocular disorders, such as the development of Convolutional Neural Networks (CNNs) from scratch, the training of Recurrent Neural Networks (RNNs), and the improvement of BERT-based models.

Furthermore, there are notable parallels between the previously cited study and our investigation. As in their examination, our study includes both the fine-tuning of a BERT architecture and the training of a Spacy-based NER model from scratch. Our goal is to experiment with their methods to improve ocular illness diagnosis and registry building, for skin disease tagging, taking advantage of the efficiency of these model architectures. We add to the larger discussion on NER approaches in the medical domains by using this comparative perspective, which highlights the potential for innovation and progress in healthcare informatics.

While building NER models for disease tagging or registry development isn't the main emphasis of another study [2], the research highlights the potential of Large Language Modeling (LLM) models in customized healthcare applications. In the paper, a system called Health-LLM is presented. It is intended to make use of cutting-edge methods such as Retrieval-Augmented Generation using personal health records (PHRs) and other relevant data. Based on a BERT architecture, Health-LLM has been refined on several electronic health records (EHR) datasets to forecast personal risks from these records.

Additionally, by adding new samples to datasets and concentrating on uncommon skin conditions, our project seeks to supplement previous work. To ensure a wider coverage of diseases and increase the effectiveness of personalized healthcare systems in precisely predicting individual risks and offering customized medical insights, this augmentation attempts to improve the comprehensiveness of datasets utilized by models such as Health-LLM. We anticipate that our approach, which can produce annotated datasets for skin disorders from unstructured records, will play a major role in expanding the variety of diseases that are represented in these models. Furthermore, additional noteworthy Language-Modeling models that are becoming more and more well-known in the healthcare industry are BioBERT[3] and BioMistral models[4] which is open sources as well. The recent addition, the open-source Llama3-OpenBioLLM-70B model based on finetuned Llama3, performs better than GPT4 for medical language modeling. Such efforts present more opportunities for the advancement of personalized healthcare applications by open-source communities. And we believe our work will support data mining for the development of such models

3. Dataset

Two main data sources were used in the construction and training of SkinNER: the American Academy of Dermatology (AAD) and PubMed.

AAD (American Academy of Dermatology). The AAD website (aad.org) provides a comprehensive repository of skin disease names, encompassing a diverse spectrum of dermatological conditions. Leveraging this resource facilitated the compilation of an extensive list of disease entities for training and evaluation purposes within SkinNER. The inclusion of rare and less-documented conditions from the AADA dataset enriched the diversity of entities recognized by SkinNER, ensuring its robustness in identifying a wide array of dermatological disorders.

PubMed. A valuable source of textual data samples for training and optimizing the SkinNER models is the well-regarded biomedical literature resource PubMed. Using a curated list of skin condition names from the AAD website catalog, we ran a PubMed search and created a corpus of text data on dermatology that included research articles, clinical narratives, and other pertinent sources. This corpus helps the SkinNER models learn from real-world instances of dermatological language and context, in addition to making training easier.

To sum up, the integration of AAD for names of skin diseases and PubMed for textual data samples yielded an extensive and varied dataset for training and assessing SkinNER. This dataset was essential in the creation of a reliable and adaptable NER system designed especially for the identification of illnesses related to the skin.

We were able to hand-compile the list of 94 skin conditions as well as conditions affecting the skin and hair

from the AAD website catalog. And we used PubMed to find the text corpus associated with these illnesses. This required searching through a huge database of biomedical literature to find pertinent papers. Fortunately, we were able to use the Entrez API to effectively interface with PubMed's resources thanks to the Biopython package. We were able to search PubMed for journals relevant to our target disorders and retrieve abstracts from these articles by utilizing this effective application. Through this procedure, we were able to compile a large corpus of textual data that included a variety of clinical narratives and research publications. This corpus proved to be invaluable for training our SkinNER models.

3.1 Preparing training data

The following steps give an overview of how we prepared our training data from AAD and Pubmed.

Step 1: Disease List creation. The skin-related disease list was obtained from the AAD website.

Step 2: Gathering documents related to the diseases. Using the PubMed Entrez API, we searched the database for papers on the list of skin conditions that had been compiled.

Step 3: Retrieving Abstracts from PubMed: Following that, the Entrez API was used to extract abstracts of the requested papers from PubMed. Research paper abstracts, which are succinct summaries, are excellent resources for knowledge about skin conditions.

Step 4: Identifying Sentences with Matching Disease Terms. We used the Spacy library to parse the extracted abstracts and find sentences in the list that had the disease terms that matched. A mapping from each identified sentence to the appropriate entities was made. The start index, finish index, and label of each entity within the sentence were included in each entity mapping. We also removed duplicate entries of sentences and combined their entities so that we could have samples with sentences with more than one entity.

Step 5: Converting to Training Formats. Finally, the created mappings were transformed into formats that worked well for both BERT-based model optimization and Spacy model training. The mappings were structured into Spacy's training data structure for the Spacy model, and the data was tokenized and formatted appropriately for BERT fine-tuning.

The resultant data contains 418,338 samples—mappings between sentences and the entities they belong to. These entities comprise cases of several skin illnesses that are sourced from the exhaustive list of 94 conditions provided by the AAD. Every sample in the dataset represents one or more of these skin conditions, giving researchers and practitioners a wide and varied corpus to work with while training and optimizing SkinNER.

We included new text corpora with examples of these skin conditions in our dataset to enhance its quality and enable thorough testing of our models. To be more precise, we used the benchmark datasets commonly used for NER models in the Medical domain. We tested models with BC5CDR, JNLPBA, and NCBI_dataset samples that have skin disease as part of context,, to check the models performance in tagging the skin disease given the context different from the pubmed dataset. This additional data, which is not part of our main training set, was only used to assess our models' performance and generalization skills. Our objective in including these varied datasets was to guarantee the resilience and efficiency of SkinNER when used in real-world scenarios outside of its original training context.

4. Models

This section outlines the architecture and techniques used to create the SkinNER system. We first introduce the first model, which makes use of an integrated Spacy pipeline with pre-existing models.

4.1. SpaCy pipeline with Token2Vec and NER

The first part of SkinNER makes use of Spacy, a well-known library for tasks related to natural language processing. We built a pipeline in Spacy that included pre-trained models for tokenization, and Entity Recognition. The NER component that was trained especially for the identification of skin-related items was added to this pipeline.

4.1.1. Architecture

The Spacy pipeline model is built with sequential processing steps, each of which is intended to carry out a particular linguistic analysis task. In our pipeline, we are using two steps, Token2Vector and NER.

In the Token2Vector step, the text is tokenized into individual words or tokens and each token is then converted into a vector. So, for each given sentence we will have a vector at the end of Token2Vector step, which will be used for the NER. NER is then performed to detect and categorize named things, in our case Skin diseases. The output of this NER will be the list of entities found by the pipeline for the given sentence. We generated the model configuration using the tool provided by the pipeline as part of their "Quickstart" section.

4.1.2. Training

Our training process involved creating a competent Named Entity Recognition (NER) model for SkinNER using the Spacy train command from the Spacy library. The PubMed dataset was first divided into training and testing subsets, using an 8:0 ratio to guarantee comprehensive assessment and validation of our model's functionality.

We carefully set up our model's parameters for the best results during the training phase. We trained the pipeline using a batch size of 8. In addition, we limited the number of steps in the training to 6000 and used a patience parameter of 1400 to control when to stop. Thus, to ensure an efficient and effective training procedure without overfitting, training was stopped at about 4200 steps when no discernible performance increases were seen. We used the specific "*spacy.Tok2Vec.v2*" model for the Token2Vector model for vectorization and "*spacy.TransitionBasedParser.v2*" for the NER model part of the pipeline.

4.1.3. Evaluation methods

We employed the original test set split from the PubMed dataset for evaluation. Using Spacy's benchmarking accuracy feature, we thoroughly evaluated our best model to get important classification assessment metrics, such as F1-score, precision, and recall.

4.2. Fine-tuning BERT

To further expand the potential of SkinNER, we set out to use cutting-edge deep learning methods—more precisely, fine-tuning pre-trained BERT models—to improve Named Entity Recognition (NER) performance for entities connected to the skin.

4.2.1. Architecture

The transformed-based architecture of BERT (Bidirectional Encoder Representations from Transformers), a cutting-edge natural language processing (NLP) model created by Google AI, is essential to our BERT-based methodology. The bidirectional nature of BERT's architectural design sets it apart from other systems and enables it to extract contextual cues from words that come before and after a text sequence. Because of its bi-directionality, BERT may produce complex semantic representations of words and phrases, which makes it easier to comprehend the underlying semantics of textual material.

We used a variation called "bert-base-uncased" from HuggingFace which has 12 transformer layers and 768 hidden units, making it very flexible for a wide range of NLP applications. To perform sequential transformations on the input text and gradually extract abstract information at each layer, these transformer layers work together cohesively. At each transformer layer, the model can encode contextual information and word dependencies by using attention processes to focus on important parts of the input sequence.

The pre-trained BERT model's parameters are adjusted during the fine-tuning phase to better fit our unique task of Named Entity Recognition (NER) for entities connected to the skin. To modify the model, its weights are recalculated using task-specific annotations, and its performance is improved by iterative updates using gradient descent and backpropagation. Through the pre-trained BERT model's fine-tuning of our well-annotated dataset, we make use of its ability to recognize complex linguistic patterns and contextual cues relevant to dermatological text data, hence improving its recognition and classification of skin-related elements.

4.2.2. Training

We carefully designed a configuration in our training regimen for the BERT-based model to maximize learning dynamics and model resilience. We adjusted crucial factors controlling the training procedure by utilizing HuggingFace's TrainingArguments. We used an epoch-based evaluation technique to make sure that performance was consistently assessed at the end of each training period.

The learning rate, which we set at 2e-5 to achieve a compromise between fast convergence and stability, was a crucial component of our training setup. By using a 16-device batch size for both training and evaluation, we were able to effectively handle memory constraints and expedite training iterations using parallel processing. The necessity for thorough learning from the dataset, which ensures the model's exposure to a variety of cases for robust performance, led to the decision to train for three epochs.

In addition, we included weight decay with a 0.01 coefficient to reduce overfitting and promote model generalization. The training pipeline was made more efficient by utilizing HuggingFace's trainer object, which handled data loading, optimization, and model evaluation with ease.

4.2.3. Evaluation methods

Straight from our refined BERT model's training process, the HuggingFace Trainer gives users instant access to critical metrics. These metrics provide important information about the model's development and performance. They include precision, recall, F1-score, and losses from training and validation.

5. Evaluation metrics

For a more comprehensive evaluation, we employ the Precision, Recall, and F1 scores.

Precision: Precision measures the accuracy of positive predictions made by the model. It is the ratio of true positive predictions to the total number of positive predictions made by the model.

Precision = True Positives / (True Positives + False Positives)

Recall: Recall measures the ability of the model to correctly identify all positive instances in the dataset. It is the ratio of true positive predictions to the total number of actual positive instances in the dataset.

Recall = True Positives / (True Positives + False Negatives)

F1 score: The F1-score is the harmonic mean of precision and recall. It provides a single score that balances both precision and recall.

F1-score = 2*(precision*Recall) / precision+recall

6. Results

Model	Precision	Recall	F1 score
Bert- Fine Tuned*	**98.222%**	**98.51%**	**98.366%**
sPaCy-NER	86.9%	98.69%	92.42%

Table 1: Evaluation results for both models

In Named Entity Recognition (NER) for skin-related entities, the improved BERT model performs better than the Spacy-NER model, with precision, recall, and F1-score values of 98.222%, 98.51%, and 98.366%, respectively. On the other hand, the Spacy-NER model has a lesser precision of 86.9%, albeit with a respectable recall and F1-score of 92.42% and 98.69%.

These findings demonstrate the BERT-based approach's better performance and accuracy, underscoring its usefulness in precisely locating skin-related entities in textual data and developing dermatological informatics applications.

The spacy model is tested with a benchmark dataset and performed with 65 percent accuracy. The main reason for the reduction in performance could be attributed to the change in the form of context in Pubmed compared to the benchmark dataset which has context in the clinical setting.

We can also see some sample outputs predicted from the Spacy pipeline below, visualized using Spacy's Displacy module.

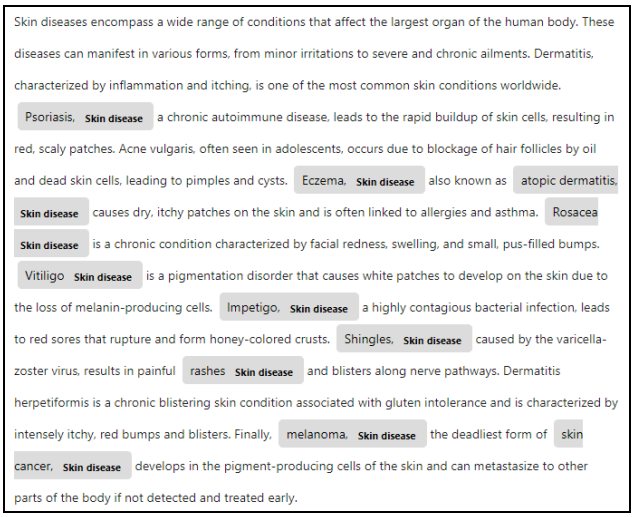

Fig 1. Sample output from using the trained spacy NER pipeline.

The sample above demonstrates the model's capability by demonstrating its ability to detect the majority of skin illnesses in the provided text. Based on context, it generalizes to tag more recent skin conditions like "atopic dermatitis," although it still ignores certain diseases showcasing the potential for further improvement.

7. Conclusion

In this paper, we presented a NER solution called SkinNER, with two models of Named Entity Recognition (NER) for skin diseases. We showed an emphasis on contrasting the Spacy-NER approach with a refined BERT model. We exhibit the enhanced performance of the refined BERT model using thorough experimentation and evaluation, attaining notable values for precision, recall, and F1-score over the spacy pipeline.

These results highlight the value of using cutting-edge deep learning strategies to improve the precision and effectiveness of NER tasks in dermatological informatics, such as fine-tuning pre-trained BERT models. The BERT-based approach's high precision and balanced recall indicate its potential for clinical decision support and research applications, opening new avenues for dermatological diagnosis, treatment, and information acquisition. We plan to finetune other LLM models with more parameters to further test this use-case scenario and see if it improves the performance of this finetuned BERT-based model.

8. Github link

https://github.com/rakesh09111996/NER_Skin-disease

References

[1] Jin, M., Yu, Q., Shu, D., Zhang, C., Fan, L., Hua, W., Zhu, S., Meng, Y., Wang, Z., Du, M., & Zhang, Y. (2024, February 1). Health-LLM: Personalized Retrieval-Augmented Disease Prediction System. arXiv.org. https://arxiv.org/abs/2402.00746.

[2] Lee, Jinhyuk, 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, no. 4 (2020): 1234-1240.

[3] Labrak, Yanis, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains." *arXiv preprint arXiv:2402.10373* (2024).

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[5] Liu, Zihan, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. "NER-BERT: a pre-trained model for low-resource entity tagging." arXiv preprint arXiv:2112.00405 (2021).

[6] Bhasuran, Balu, and Jeyakumar Natarajan. "Automatic extraction of gene-disease associations from literature using joint ensemble learning." PloS one 13, no. 7 (2018): e0200699.

[7] Doğan, Rezarta Islamaj, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: a resource for disease name recognition and concept normalization." Journal of biomedical informatics 47 (2014): 1-10.

[8] Habibi, Maryam, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. "Deep learning with word embeddings improves biomedical named entity recognition." Bioinformatics 33, no. 14 (2017): i37-i48.

[9] Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), pages 73–78, Geneva, Switzerland. COLING.

[10] Lou, Yinxia, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. "A transition-based joint model for disease named entity recognition and normalization." Bioinformatics 33, no. 15 (2017): 2363-2371

[11] McCann,B. et al. (2017) Learned in translation: contextualized word vectors. In: Guyon,I. et al. (eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp. 6294–6305. http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf.

[12] Sachan, Devendra Singh, Pengtao Xie, Mrinmaya Sachan, and Eric P. Xing. "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition." In Machine learning for healthcare conference, pp. 383-402. PMLR, 2018.

[13] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[14] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).