Bio-NER tool for symptom identification in preliminary diagnosis

Chelsy Fernandes and Megha Manoj
BITS Pilani Dubai Campus, United Arab Emirates
f20200003@ dubai.bits-pilani.ac.in; f20200016@dubai.bits-pilani.ac.in

Hypothesis

Propose a tool which acts as a middleware in identifying symptoms from the patient's conversation with medical symptom-based diagnostic chatbots- both traditional and voice-based. The research assumes that the mode of input is text or speech-to-text communication between the chatbot and patient. As individuals engage in a casual tone of voice due to the informal setting of a conversation with a bot, their input on the problem they face is very vague. The tool shall be used to narrow down the symptoms which can be utilized as a term-based summary for diagnosing diseases.

Abstract

The implementation of text mining within the medical field has been rapidly changing over the years. While a plethora of use cases of the niche exists, there has been a huge concern with the accuracy of the models built. The issue can mainly be attributed to the ambiguous interpretation of datasets and inputs by the models. In order to provide a plausible solution to the problem, we conducted a literature survey on the past research done within the field of AI to understand the causes of the concern within a small area of the medical sector- disease symptom diagnosis. Based on the survey carried out, we intend to create a middleware tool that can identify general symptoms from chatbot conversations which shall act as a narrowed down input to boost the performance of medical bots. The BERT for sequence-classification model is used in this paper.

Introduction

Natural Language Processing (NLP) is a sub-field of AI which allows the machines to communicate in the language as humans do. NLP techniques can be used for a variety of tasks like classification of text, sentiment analysis and named entity recognition and many more. NLP has transformed the way people are interacting with technology and enhanced the capabilities of data. Machine translation has become easier because of the utilization of NLP. Due to the inclined progress of the techniques as well as the use of transformers for Large Language Models (LLMs) within the healthcare sector, it has contributed to a majority of use cases in medical informatics. Hospitals around the world today use Electronic health records often referred to as EHR which replaces the traditional pen and paper method. Out of the commonly implemented techniques, biomedical text mining plays the most important role of extracting data from medical texts, patients records etc. The main aim of this field is to present valuable outcomes from large unstructured data. The data generated from this can be used for prediction of diseases and used for decision making purposes.

The utilization of Natural Language Processing within the medical field is constantly evolving. It is used for a diverse range of purposes- from patient information summarization to assistant chatbots. At present, its use cases in medical informatics with the availability of large datasets have aided in enhancing the accuracy of medical NLP applications. A major use of medical NLP in the current scenario involves attempting to deploy medical chatbots with high accuracy. However, due to the persistence of ambiguity within these bots, the risk of real-time deployment is extremely high, especially with regards to incorrect diagnosis that can result in grave ethical and legal concerns. Since the root cause of ambiguity is the vague description of symptoms given by patients, we intend to extract the exact symptoms from their conversation and utilize named-entity recognition (NER) to increase the accuracy of the diagnosis. NER is used for extracting information which involves the processing of structured as well as unstructured documents and identification of expressions which refer to places, people, companies, organizations etc. NER is a primary task and forms the basis of NLP systems. It includes two tasks, first one is identifying names in text and secondly classification of the names into some categories like person, places etc [1].

BioNER which is Biomedical Named Entity Recognition is an identification task for the identifying of chemicals, proteins, genes, viruses, DNA and so on. BioNER uses Supervised Learning. Features play a vital role in improving the efficiency of the recognition process. Features are distinguishing abilities or characteristics that help in identifying the occurrence of the entities.

For this purpose, we take a dataset consisting of recorded patient-doctor conversations that mirror the typical conversation an individual may have with a chatbot and train a BERT transformer for NER.

Literature review

There has been a prolonged debate on the credibility of AI applications within the medical sector. In order to do so Julia et al [2] focus on the clinical assistance offered to doctors by support systems (CDSS) in tending to ER patients suffering from cardiac arrest by considering an array of factors which influence the system's decision making process. By focusing on the explainability of CDSS, they are assured that transparency of AI systems can be achieved which would resolve the issue of "Black-box" AI which does not offer much information on its working. An additional factor which can contribute to the solution is detailed documentation of the program.

The creation of a medical domain specific BERT model termed the BioBERT [3] was done in order to accommodate the prolific range of biomedical text which up until 2019 had restricted applicability of general word representation models due to variations between the actual datasets and that used within the field. Additionally the incapability of models such as BERT to perform on larger corpus did not improve the situation. BioBERT offers a standard solution to the problem by pretraining on large documents such as PubMed with evaluation performed by testing its productivity when tasked to do 3 different text mining methods- NER, question answering(QA) and relation extraction(RE). The authors suggest expanding the testing on other text mining tasks to validate the model's range of application.

ComBot [4] is a 3 bot model -consisting of a task bot, follow up bot and an empathy bot. The first model gathers information from a questionnaire and presents the user with questions related to the user input. The data provided is then processed to output sentimental feedback to the user. It offers insights into the application of dialog models in medicine. While the model performs optimally, it can falsely consider chats with a stressed individual to be a false positive and furthermore, does not understand when a conversation has concluded, which may induce agitation on the human-side of the conversation as they feel forced to continue chatting.

The study [5] focuses on comparing the implemented logic of CDS using ChatGPT with that of suggestions made by humans. It was observed that the AI model gave out-of-the-box suggestions which seemed better in comparison to that made by the latter upon evaluation by clinical professionals on the basis of several factors. The model had a low acceptance rate and was found to be moderately applicable- which calls for optimization of generated recommendations by Large Language Models and identification of potential areas of improvement of the model to develop in-depth applications.

The paper[6] gives an overview of the concerns surrounding utilization of AI in medical support systems with respect to ethics, legality and society. It elaborates upon the various factors AI models should consider- algorithmic transparency, patient safety, regulation, accountability, liability and the impact on the patient-physician relationship. With respect to these factors the authors express an inclined position towards adoption of a “Ethics by Design" approach to lessen the stress on these concerns with an awareness of the potential of AI in healthcare with an emphasis on the issues of complex ethics, legality, and social challenges that need to be addressed.

The research paper [7] analyses named entity recognition (NER) in the medical field using NLP and the methodologies proposed with respect to its role as a preprocessing tool. They derived data from various research paper sites and analyzed the different terms extracted from the texts along with their synonyms and other relevant terms. The methodology adopted by the authors involves devising regular expressions for finding relevant papers and verifying whether the papers are of use by checking its title and number of references. Then they looked into the names of the sections out of which they were able to conclude that ML models were used by the majority of the conducted research. Through section-wise, NE analysis they were able to draw conclusions on what type of concepts were dominant within the field.

The research [8] compares the performance of popular AI chatbots- Microsoft Bing AI Chat, Google Bard and OpenAI’s ChatGPT in classifying disease symptoms obtained from the subreddit platform. The aim of this comparison was to determine the sense of urgency in medical cases as perceived by the chat systems. Out of the 3 models, Google Bard’s beta version showed high levels of accuracy after a cross-review conducted with experts in the field of emergency department. However, all models exhibited similar levels of low performance when it was tasked with classifying actual emergencies. The same conclusion was deduced for non-emergency situations which fell into the ‘cannot be determined’ category. The authors also listed the vague responses of these chatbots in giving a direct answer on recommending to visit the emergency department as a limitation occurring from the general training data supplied to them. Another limitation is the lack of sentiment in understanding the information stated by the patient which results in misinterpretation of how intense the case is or assume it to be more severe than it actually is.

The authors [9]discuss the low cognitive ability of LLMs despite their present success and a new method of prompt engineering called CARP introduced by them. The technique takes advantage of the models’s generalization feature and a topic specific labeled dataset by implementing a fine-tuned model using kNN in the in-context search. It carries out the process in 3 steps- searching for clues relevant to semantics and keywords, using the found clues and the input text to influence a diagnostic reasoning process and generate a label for the text. CARP overcomes the limited token use of LLMs through the supervised in-context search to yield better results with low resource setups.

The study[10] focuses on the use of intelligent diagnostic systems to assist health workers in making adequate decisions and offer recommendations about treatments to patients. With respect to this aim, the paper compares performance of different traditional models, LLMs and NLP models. The comparisons revealed a faster diagnosis in machine learning models but required large amounts of reviewed data for training while LLMs gave a sensitive response to the input texts due to their pretrained nature. The performance of LLMs, ChatGPT and BARD were much better in comparison to that of the NLP models which failed to provide precise diagnosis and required further consultation to reach a decision. They also mention the ambiguity and irrelevant nature of the response as a downside of this application which could be resolved by inventing new language models.

The authors [11] point out the single task utility of text based data within the medical field stating it as a failure. According to them the importance of safe prediction is yet to be attained through a discreet framework. In order to build such a framework, one must go beyond the standard dataset and use expert reviewed data and make use of evaluation metrics other than the BLEU score which are capable of assessing LLMs on the basis of potential harm, precision, reasoning among other levels of expertise. They have proposed a benchmark combining six different medical question answering datasets called MultiMedQA as well as a new dataset called HealthSearchQA which contains data related to medical queries searched online. The authors evaluated PaLM and Flan-PaLM models using MultiMedQA out of which the Flan-PaLM model achieved high accuracy using the dataset. However the model had several gaps in responding to consumer queries with a rate of 29.7% to generate harmful outputs. To counter this, a Med-PaLM model was developed.

The paper [12] discusses the usage of NLP methods for the development of health diagnosis conversational systems. In this research paper a chatbot had developed a telehealth system on the basis of rules of fuzzy logic and fuzzy interference. The main objective of the service is to evaluate the tropical disease symptoms in Nigeria. The system and the chatbot were connected via the Telegram Bot API, and a Twilio API was utilized to establish connection between the system and the subscriber of the SMS service. The service uses the knowledge of the information obtained from the medical ontologies about the identified illnesses and symptoms. Based on the entered symptoms, a fuzzy support vector machine (SVM) is utilized to accurately forecast the illness. NLP recognizes user inputs, which are then sent to CUDoctor for decision support. The user is then issued a notification message indicating that the diagnosing process has come to a conclusion. The end product is a medical diagnosis system that uses user self-input to provide a personalized diagnosis in order to accurately diagnose illnesses. The system usability scale (SUS) was used to assess the produced system's usability; the result was an overall positive rating with a mean SUS score of 80.4.

The study[13] discusses the usage of deep learning for the development of a question answer model for the healthcare sector. It fills the gap of the unavailability of answers to medical questions and provides for the development of different training models which include LSTM, Embeddings etc. The research paper does a comparative study of the different training models and the evaluation is done on the basis of BLEU score. The paper also addresses the issues related to BLEU score. The paper trains an end-to-end network for answering the medical questions and proposes a model for accurate results.

Computer Aided Diagnosis (CAD) [14] represents a rapidly advancing and diverse field in medical analysis.Recent years have witnessed significant tries aimed at creating applications for computer-aided diagnostics,driven by the recognition that flaws in medical diagnostic processes can lead to profoundly misleading treatment decisions. Machine Learning(ML) plays a crucial role in Computer Aided Diagnostic tests .Simple equations prove insufficient for accurately identifying objects like body organs , underscoring the necessity for pattern recognition training based on the instances. In the biomedical domain, the integration of pattern detection and ML holds the promise of enhancing the reliability of disease detection and approach , all while maintaining an impartial decision-making process. ML presents a robust framework for developing advanced algorithms capable of analyzing high-dimensional and multi-modal biomedical data. This survey paper searches into a comprehensive examination of various ML algorithms employed for the detection of diseases such as heart disease and diabetes , emphasizing the combination of algorithms and techniques for ML in disease detection and disease-making processes.

Prioritizing a healthy lifestyle is a key focus for individuals to ensure a joyful life, it is crucial to consistently monitor one’s health. The most reliable method for doing so involves undergoing regular health checkups and obtaining disease diagnoses. While numerous treatments are available for various illnesses, there is a lack of a centralized resource providing comprehensive information on diseases and their respective treatments. This research aims to address this issue by offering a solution where users can receive treatment recommendations simply by entering their symptoms. They proposed [15] approach seeks to streamline the tedious process of visiting hospitals, especially during the ongoing pandemic and securing a doctor’s appointment. This research focuses on applying various concepts of NLP and ML to develop a chatbot application. Users can engage with the chatbot much like they would with a human and they will attempt to predict the disease and suggest suitable treatment options. The envisioned system holds significant potential for individuals, enabling them to conduct routine health check-ups, and encouraging them to adopt cost-free measures for maintaining good health.

The main aim of this paper [16] is to detect Cerebrovascular disease from Electronic medical records (EMR) by utilizing NLP. The previously experimented methods had time-delays and issues with coding. The model was trained on several NLP models using supervised learning. The comparison of the model was done with International Classification of Diseases (ICD-10-CA). Results of the study was that Cerebrovascular disease dominated by 11.8%. Among the NLP models applied XGBoost combined with cTAKES along with the term-document frequency and inverse document frequency. The NLP models had a greater validity as compared to ICD codes in the detection of CeVD. The conclusion of the study was that the Natural Language Processing models could have been used for the development of an EMR tool for the identification of future CeVD cases.

The paper [17] discusses diabetes prediction using predictive analysis. The paper summarizes the applications of predictive analysis in the healthcare sector especially in predicting diabetes. Study goes in depth to discuss various kinds of models used for prediction and the currently used models for predictions. The study is conducted on a hybrid model that improves the accuracy of the prediction. The paper also mentions the gap in the existing literature and suggests developments for predicting diabetes more accurately.

This paper [18] talks about the usage of deep learning and machine learning in making predictions in the healthcare sector. The paper highlights the requirement of reliability as well as an efficient methodology for predictive analysis in healthcare. The paper discusses the different machine learning models like k-nearest neighbor, logistic regression, SVM, random forests and decision trees being used for making predictions. The algorithms were used in predicting diabetes, heart diseases etc. Deep learning models like Recurrent Convolutional Neural Network RCNN, Convolutional Neural Network (CNN) and LSTM can be used for making predictions of diseases. The deep learning techniques are used to predict heart diseases and improve the accuracy of the model.

This research [19] discusses the use of NLP for clinical notes analysis relating to chronic diseases. The main focus of the study is the circulatory system diseases like hypertension and heart disease whereas there is less focus on the nutritional and endocrine issues. Clinical notes are mostly analyzed using machine learning techniques as compared to the rule-based approaches. The study focuses on disease phenotype classification along with addressing extraction of coexisting data from text that contains structured data. The drawbacks of this research are the limited availability of data for training purposes in the medical domain. Suggestions are given in this paper for further use of NLP methods for extortion of clinical data in the future.

The paper [20] discusses the machine learning algorithms used in diagnosing liver diseases and its prediction. The paper does a comparative study between various algorithms like Artificial Neural Network, SVM, Convolutional Neural Network (CNN), Naïve Bayes and Logistic Regression. The results of the study highlights that the highest accuracy Is achieved by CNN. The paper further focuses on the use of scan images for liver disease predictions. The dataset used for analysis is pre-processed and machine learning algorithms are applied on it. Evaluation metrics are generated from the various algorithms used. The study concluded that CNN displayed a higher performance.

The study [21] puts forth a hybrid dependency parser for the extraction of data from biomedical texts. The limitations of the previously existing parsers were identified from the semantic relationship between extraction of models and sentence constituents. The paper states the use of semantic relationships in the co-occurrences in the bio-medica texts. It devises a count system based on the frequency of co-occurrences and the semantic relationships. It is used in the predictions of protein functions and gene diseases. The system implements LEAP-FS, Text KNN, Prodisen for predictions. The paper also discusses about the extraction of data using the existing dependencies and focuses on the prediction pf protein functions and interactions. The results show that EDC_EDC performers better as compared to other systems.

A study [23] conducted on extraction of general symptoms from medical files makes use of a SVM combined with stochastic gradient descent to create a classification model. The dataset used within the research included 59,412 symptoms which were annotated by humans to ensure correct labeling. The model has a F1-score of 80% with the false positive rate being 60%, which is its major drawback. The authors state that the disadvantage mainly comes from the grammatical aspect of symptoms denoted as a modifier as well as an action in the data as well as the lack of variation in mentions of the symptoms.

The paper discusses [24] the fundamentals of NER and the preprocessing of the data, feature processing and evaluation. The aim of this research is to classify the genes, proteins into their names, chemical names and the names of the disorders associated with them. The gene naming focuses on two main categories GENETAG and JNLPBA. The F1 score of the model implemented is greater than 80%. For the names of the disorders Arizona Disease and Corpus and Bio-text are used. The Arizona Disease and Corpus resulted in f1 score 81.8% whereas biotext had F1-score 54.84%. For the chemical names ChemSpot gave a F1 score of 85.6%.

The paper [25] proposes to improve named entity recognition with the help of deep learning and word embedding. LSTM CRF is a method introduced in this paper; an independent named entity recognition methodology based on deep learning word embedding. The LSTM CRF has a better F1 score compared to the general CRF method. LSTM CRF performs the best on 28 evaluations out of 33. It covers 5 different types of entities- chemicals, species, diseases etc.

Dataset Description

With respect to the limitations of paper [23], we chose a dataset from Kaggle and it contains 6662 recordings of patients describing their symptoms. It has 13 features- patient audio files, the transcribed text related to them, background noise involved and its audibility, audibility of the speaker, quality of recording, audio clipping, audio file name, audio file download and speaker id. The audio files account for about 8 hours of symptom description. Its main purpose is to train conversational agents in the medical sector. The dataset consists of 25 labels related to common symptoms experienced by sick people such as cough, back pain, etc and a variety of phrases which express the precise symptom they wish to convey with respect to the transcription.

K	L
when i raise my hand i feel pain in my shoulder.	Shoulder pain
I have acne all over my face	Acne
I get clusters of pimples on my face that never go away.	Acne
My foot hurts me a lot of playing football	Foot ache
I'm having a hard time reading because the letters are all fuzzy	Blurry vision
i was playing football and injuried with joint pain.	Joint pain
I feel discomfort throughout the body in general	Body feels weak
there pain in my foot	Foot ache
I don't know why I'm constantly sad.	Emotional pain
I got a divorce last year and I just can't stop dwelling on how to get revenge on my ex hus	Emotional pain
Chronic disease of hair follicles and sebaceous glands	Acne
My chest hurts when I smoke	Internal pain
I feel lightheaded when I stand up	Feeling dizzy

Figure I. Dataset

Methodology

I. Data Preprocessing

Since our focus is on extraction of symptoms from the transcribed text, we utilize only 2 features of the dataset- patient phrases and the respective prompts. In order to correct the irregularities present within the dataset a train-test split is performed on the data. The data is split into train, test and validation. 70% train, 20% test and 10% validation set. We tokenized the sentences and added padding and truncation to create consistent records of the same word count. To format the inputs to the models, attention masks are created for differentiating the real tokens from pad. Additional numeric conversion was performed by encoding the labels to numbers. Since the model is a transformer,the split and preprocessed datasets are converted to tensor data loaders for training, testing and validation.

II. Training and Testing Model

In correspondence to the literature review conducted, we built a small BERT model that performs extraction of general symptoms on the basis of the dataset and its variant descriptions of the symptoms. Bert utilizes a multi-layer transformer that is a bidirectional encoder for representing the given text in a high-dimensional space. It takes in the entire context of each of the words in a sentence. This makes it easier for the model to understand the text. The model was constructed on Colab under the T4 GPU environment for training and testing. It was previously trained on a large corpus in addition to the training we incorporate with our chosen dataset. For the purpose of classification, we added an extra linear layer which undergoes training along with the pretrained model on textual inputs. By supplying the preprocessed data to the model with an AdamW optimizer and a learning rate adjuster for modifying the rate during training. Once all training and testing is completed, the model is saved for linking it to a streamlit application which shall demonstrate the input and outputs of the tool. The figures II and III depict the demonstration of the web app.

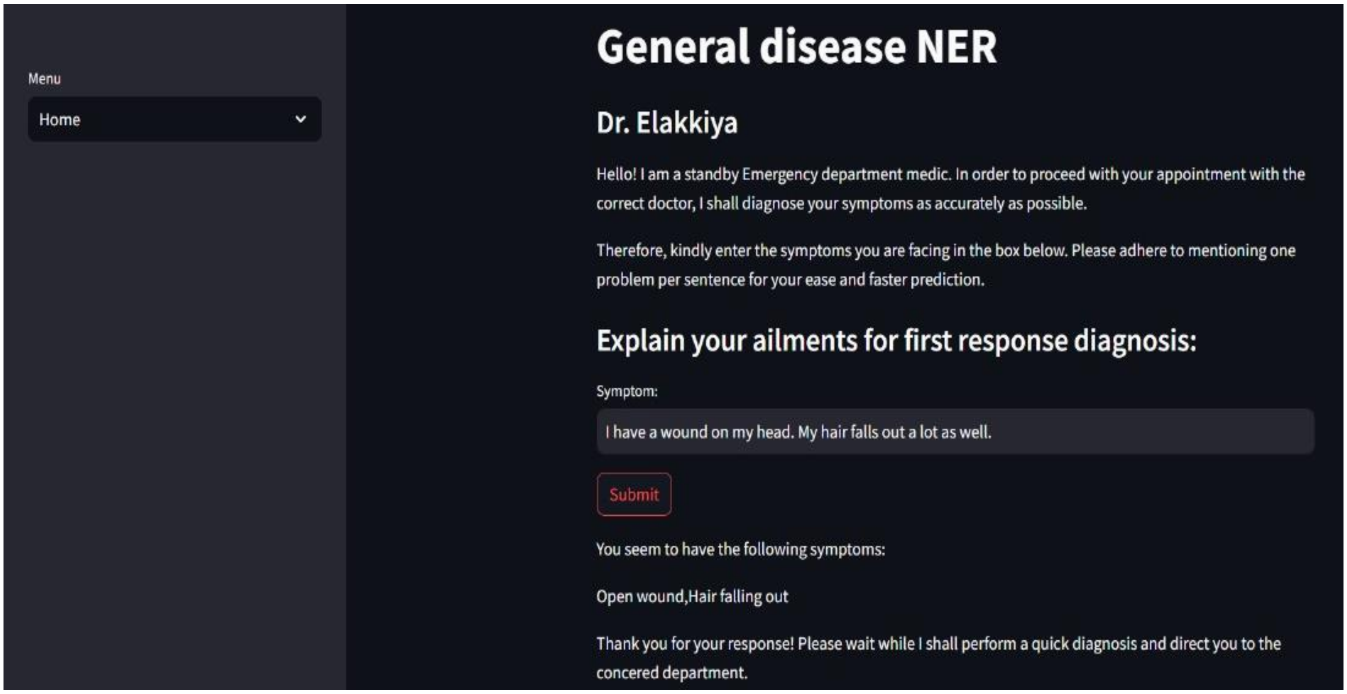

Figure II. Home page of web app

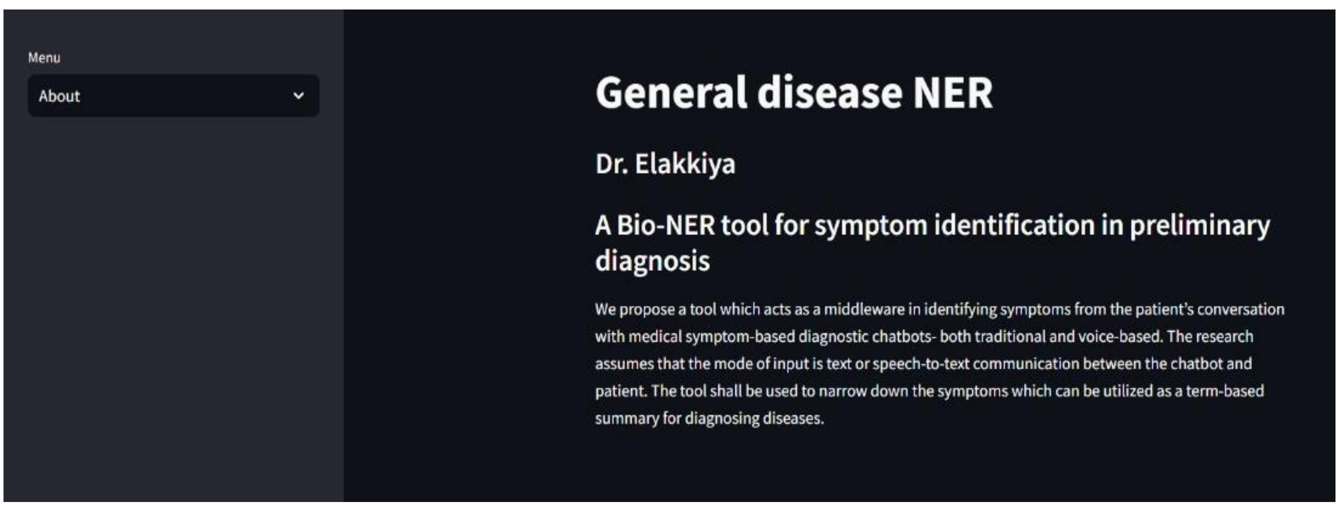

Figure III. About page of Web app

Results and Discussion

Table I displays the performance of the model on train, test and validation datasets. The BERT model uses the AdamW optimizer and it runs for 4 epochs. The table below shows the accuracy and loss at each stage in training respectively.

Epoch no.	Training accuracy	Training loss
1	0.37	2.61
2	0.93	1.01
3	0.99	0.36
4	1.00	0.19

Table I. Evaluation metrics for model training

Upon calculating the scores for the evaluation metrics- accuracy and loss of both validation and testing we received a loss of 14% and an accuracy of 99.4%. While keeping in mind that the performance of BERT varies with respect to the dataset it works with, through testing a variety of unseen inputs, we concluded that the performance of the model is in par with its previous implementations and it can offer improved results with expansion in the vocabulary used for denoting the labels of the symptoms. When compared with the SVM classification algorithm implemented in paper [22], the false positives within the web app is comparatively less as it gives a much more accurate result.

In terms of limitations of the model, it assumes that if an empty string is provided as input, the patient is in emotional pain, which is a bit of an ambiguous take. However , since there is no proper contextual conversation taking place in the tool, it is difficult whether this limitation is a valid one. Additionally, the dataset used lacks a variety of symptoms which can help enhance the level of symptom identification. For future work, using an expert-approved dataset obtained from hospitals can help fine-tune the performance of the tool.

Conclusion

Through the paper, we have attempted to address the research gap related to the ambiguous nature spotted in disease prediction due to incorrect identification of symptoms in medical chatbots. We propose the creation of a layer between the inputs and the disease diagnosis model to generate a list of symptoms that the patient experiences. These recognised signs of diseases can then be inputted to a disease prediction model. By using a labeled dataset which contains conversational phrases of a patient describing their symptoms, we create a web app that intakes the symptoms described by the user and performs named entity recognition to derive the symptoms as a demonstration of the tool.

The built model has an accuracy of approximately 94% and performs at a highly precise level. However, due to the limited number of symptoms in the dataset, its usage is limited. Furthermore, in case any typos are present within the input, it may hinder the recognition of the keywords which the model searches for. For future works, research to expand the data used in training and testing, as well as deploying the second part of the model- for disease prediction. It would also prove beneficial to incorporate automatic word correction related to medical terms. Since the complexity level of the vocabulary used to describe the symptoms is low, addition of more complex labels or multiple labels for the same phrases may add to overall development of the model at a higher level.

References

1. Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat, “Named Entity Recognition Approaches ”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008.

2. Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, Gilbert TK, Hagendorff T, Holm S, Livne M, Spezzatti A, Strümke I, Zicari RV, Madai VI; Z-Inspection initiative. “To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems.” 2022 Feb 17 doi: 10.1371/journal.pdig.0000016.

3. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, Volume 36, Issue 4, February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682.

4. Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2021. Gathering Information and Engaging the User ComBot: A Task-Based, Serendipitous Dialog Model for Patient-Doctor Interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, Online. Association for Computational Linguistics.

5. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, McCoy AB, Sittig DF, Wright A. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. medRxiv [Preprint]. 2023 Feb 23:2023.02.21.23286254. doi: 10.1101/2023.02.21.23286254. PMID: 36865144; PMCID: PMC9980251.

6. Anto Čartolovni, Ana Tomičić, Elvira Lazić Mosler, Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review, International Journal of Medical Informatics, Volume 161, 2022, https://doi.org/10.1016/j.ijmedinf.2022.104738.

7. Goulart, Rodrigo Rafael Villarreal,Strube de Lima, Vera Lúcia,Xavier, Clarissa Castellã, “A systematic review of named entity recognition in biomedical texts.” 2011 June 01, Journal of the Brazilian Computer Society, Vol 17 page 103-116.

8. G. Zúñiga Salazar, D. Zúñiga, C. L. Vindel , A. M. Yoong , S.Hincapie ,A.B. Zúñiga, P.Zúñiga , E.Salazar ,B.Zúñiga, “ OpenAI’s ChatGPT, Google Bard, and Microsoft Bing AI Chat.” , 2023 September 18. doi:10.7759/cureus.45473.

9. Sun, Xiaofei et al. “Text Classification via Large Language Models.” *ArXiv* abs/2305.08377 (2023).

10. Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, Genoveffa Tortora, “Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot”,Expert Systems with Applications,Volume 235, https://doi.org/10.1016/j.eswa.2023.121186.

11. Singhal, K., Azizi, S., Tu, T. *et al.* “Large language models encode clinical knowledge.” *Nature* 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2.

12. A. Dogra, Nicholas A. I, Ndaman, Israel O., Sanjay Misra, Olusola O. Abayomi-Alli, Ro. Damaševičius, “Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic”, 2020 September 29. https://doi.org/10.1155/2020/8839524

13. Abdallah, Abdelrahman & Kasem, Mahmoud & Hamada, Mohamed & Sdeek, Shaymaa. (2020). “Automated Question-Answer Medical Model based on Deep Learning Technology”. 1-8. 10.1145/3410352.3410744.

14. P. Hamsagayathri and S. Vigneshwaran, "Symptoms Based Disease Prediction Using Machine Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 747-752, doi: 10.1109/ICICV50876.2021.9388603.

15. J. Agarwal, M. Kumar and A. K. Srivastava, "Symptoms Based Disease Diagnosis and Treatment Recommendation," 2021 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2021, pp. 162-167, doi: 10.1109/ComPE53109.2021.9751805.

16. Pan J, Zhang Z, Peters SR, Vatanpour S, Walker RL, Lee S, Martin EA, Quan H. “Cerebrovascular disease case identification in inpatient electronic medical record data using natural language processing.” Brain Inform. 2023 Sep 2;10(1):22. doi: 10.1186/s40708-023-00203-w. PMID: 37658963; PMCID: PMC10474977.

17. Jayanthi, N., Babu, B.V. & Rao, N.S. “Survey on clinical prediction models for diabetes prediction.” *J Big Data* 4, 26 (2017). https://doi.org/10.1186/s40537-017-0082-7

18. Badawy, M., Ramadan, N. & Hefny, H.A. “ Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Inf Technol”* 10, 40 (2023). https://doi.org/10.1186/s43067-023-00108-y.

19. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V, “Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review”, JMIR Med Inform 2019, doi: 10.2196/12239.

20. P. S. Harshini, K. Naresh, S. R. Pamulapati and A. Lavanya, "Diagnosis of Liver Diseases Using Machine Learning Algorithms and their Prediction Using Logistic Regression and ANN," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6, doi: 10.1109/CONIT59222.2023.10205819.

21. K. Taha, "Extracting Various Classes of Data From Biological Text Using the Concept of Existence Dependency," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 6, pp. 1918-1928, Nov. 2015, doi: 10.1109/JBHI.2015.2392786.

22. Divita G, Luo G, Tran LT, Workman TE, Gundlapalli AV, Samore MH. General Symptom Extraction from VA Electronic Medical Notes. Stud Health Technol Inform. 2017;245:356-360. PMID: 29295115.

23. https://ngrok.com/docs/using-ngrok-with/googleColab/

24. B.Alshaikhdeeb, K.Ahmad, “ Biomedical Named Entity Recognition: A Review”, Vol.6 (2016) No. 6 ISSN: 2088-5334, https://www.researchgate.net/profile/Basel-Alshaikhdeeb/publication/311917426_Biomedical_Named_Entity_Recognition_A_Review/links/58be9718a6fdcc983145779f/Biomedical-Named-Entity-Recognition-A-Review.pdf

25. Maryam Habibi, Leon Weber , Mariana Neves , David Luis Wiegandt, “Deep learning with word embeddings improves biomedical named entity recognition”, Bioinformatics, 33, 2017, i37–i48 doi: 10.1093/bioinformatics/btx228 ISMB/ECCB 2017.

26. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent

27. https://drive.google.com/drive/folders/1XSEaE1ro77l6aRagULIA7IQgFRYbv8g?usp=sharing

28. https://huggingface.co/bert-base-uncased

29. https://medicalfuturist.com/top-10-health-chatbots/

30. https://www.revechat.com/blog/healthcare-chatbots/#:~:text=Security%20Concerns&text=Undoubtedly%2C%20it%20is%20one%20of,all%20businesses%20and%20consumers%20alike.

31. https://medium.com/@Saimely/the-future-of-care-llms-chatgpt-in-healthcare-chatbots-c8731e0cda87

32. https://publichealth.berkeley.edu/students/text-analysis-for-public-health/

33. https://datlowe.cz/6-reasons-why-keywords-are-not-enough-for-text-mining/