Disease prediction using NLP techniques

Ouabiba Hamza [1*], *Professor Sniba Farah*[1]

[1]LAMIGEP/EMSI-MARRAKECH

Abstract. This paper explores the application of the T5 (Text-To-Text Transfer Transformer) model Originating from the groundbreaking "Attention Is All You Need" concept , fine-tuned on a medical dataset to predict diseases and symptoms from unstructured medical reports. By leveraging Natural Language Processing (NLP), the system offers automated analysis, enabling quicker and more accurate diagnoses based on symptoms provided by users. The fine- tuning process involved training the T5 model to adapt to the specific language and context of medical texts. The model's performance is evaluated based on its ability to detect and predict medical conditions from user inputs.

1 Introduction

The identification of diseases plays a crucial role in healthcare, as it aids in understanding the causes or conditions that result in illness, pain, dysfunction, or even death. A disease is any condition that harms a person's mental or physical well-being, leading to significant changes in their lifestyle. The scientific field dedicated to studying diseases is known as pathology. Clinical experts analyze the signs and symptoms of a disease to make precise diagnoses [1][2]. Diagnosis refers to the process of recognizing a disease based on observed symptoms and signs, which enables healthcare professionals to understand its root cause [1][5].

Recognizing disease symptoms and understanding their effects on quality of life are vital for medical practitioners. The ability to accurately identify these symptoms is crucial in shaping patient care and facilitating the development of treatments [6][8].

To obtain accurate diagnostic results quickly and at lower costs, an efficient decision support system is necessary. The classification of diseases, based on various parameters, is a complex challenge for healthcare professionals. However, artificial intelligence (AI) can significantly aid in diagnosing and managing such cases. AI, a fundamental branch of computer science, enables systems to become more intelligent by mimicking human thought processes [9]. Many AI techniques, such as deep learning, machine learning, and data mining algorithms, have been rapidly advancing and are now applied in the medical field to predict diseases and enhance patient health management [2][10]. These techniques are instrumental in improving patient care and predicting diseases [2][10]. Some of the advantages of analyzing medical data include: (a) providing patient-centered and structured information, (b) categorizing populations based on characteristics such as diagnosis or symptoms, (c) analyzing the effectiveness and impact of drugs on individuals, and (d) recognizing clinical patterns [4]. New information technologies and computational methods are being applied to enhance the processing and analysis of medical data.

A key aspect of data processing and analysis is text classification and clustering, an area that has gained significant attention in recent years [11]. These techniques are particularly useful in the analysis of health data, as numerous medical datasets in the healthcare industry, such as those related to disease characterization, can be examined using various predictive analytics approaches [12][14].

This project is a collaboration with the Mohamed 5 Foundation, which has deployed mobile medical units in rural areas where access to healthcare is limited. These units have collected extensive patient data while ensuring confidentiality. A crucial aspect of this project was the labeling of data, which was carried out using an automated script based on a model called LLAMA3. This script takes patient data and performs the labeling process. While we worked with a portion of the dataset that was not completely labeled, the labeling process was supervised By healthcare professionals to verify the accuracy of the responses. Once the model was trained, it was deployed as a web application, enabling doctors to utilize it effectively in their practice. The model chosen was t5 Model due to it's capabilities on text to text generation T5 (Text-to-Text Transfer Transformer) converts all NLP problems into a text-to-text format, both input and output are treated as text sequences. It offers an encoder- decoder structure. The encoder processes the input text and the decoder generates the output text. T5 is trained on a large corpus using a diverse range of tasks, not just summarization. This broad training makes it versatile and effective for

[*] Corresponding author: hamza.ouabiba@gmail.com

summarization, as it can understand and replicate a wide range of writing styles and content types.

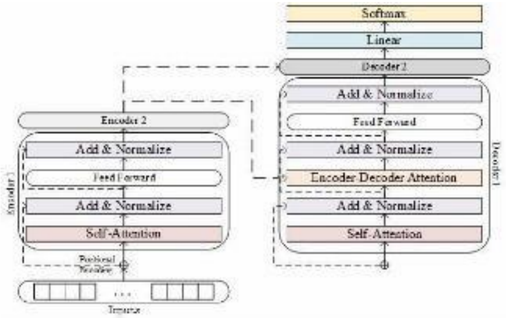

Fig. 1. T5 Model architecture

2 Ease of use

The T5 (Text-To-Text Transfer Transformer) model was introduced by Google in 2019. It revolutionized the approach to NLP tasks by framing them all as text-to-text problems, whether it’s translation, question answering, or classification. A review of the literature on the T5 model reveals its significant impact and widespread application in NLP. The model is pre-trained on a massive amount of diverse text data and then fine-tuned for specific tasks. The basic building blocks for T5 are attention-based mechanism and feedforward neural network, the high level overview of these components consist of Embedding layer, Encoder block, Encoder layer, Decoder block, Decoder layer and output layer. In which input text is tokenized and embedded into high-dimensional vectors in the Embedding layer. These embeddings represent the meaning of each token in the input sequence. Then it comes to the Encoder block which consists of a Multihead self attention mechanism and feedforward neural network. The Multihead self attention mechanism allows the model to weigh different parts of the input sequence differently while processing a particular token. It captures dependencies between words and helps the model understand the context. After the Multihead self attention mechanism the output passes through a feedforward neural network, introducing non-linearities and enabling the model to capture complex relationships in the data. The encoder is composed of multiple layers of these blocks, each layer refining the understanding of the input sequence. Similarly in decoder block it consists of Multihead self attention Mechanism, feedforward neural network along with Encoder decoder attention mechanism between them.Similar to the encoder, the decoder also uses self-attention to weigh different parts of the input sequence but in a masked manner, ensuring that each position can only attend to previous positions, while Encoder decoder attention mechanism allows the decoder to focus on relevant parts of the input sequence, helping it generate the output sequence. and again same as encoder , decoder consists of multiple layers of these blocks. The final layer is the output layer which produces an output sequence, in the case of our text summarizer this could be a condensed version of the input text. Below image explains how T5 model was the best among previous summarization models by achieving a 89.8 human score.

3 Methodology

The methodology section outlines the approach taken to collect and prepare the data, fine-tune the T5 model, and evaluate its performance. This section provides a detailed account of the steps followed in the research process.

3.1 Data Collection and Preparation

The dataset used in this study consists of medical reports collected from mobile medical units deployed in rural areas in collaboration with the Mohamed 5 Foundation. These units provide healthcare services to populations with lihospitals, gathering comprehensive patient data during diag- nosis. The reports include unstructured text describing patient symptoms, diagnoses, and other relevant medical information. The data was then annotated using an automated Python script, which attempted to identify the disease in each report using the LLAMA3 model. To enhance accuracy, the annotation process was supervised by doctors, though only for a small portion of the dataset.

3.2 Model Training and transfer learning

We employed the T5 model for this task due to its strong performance in text-to-text generation tasks. The model was fine-tuned on the annotated dataset to predict diseases based on the unstructured medical reports. The T5 model was trained and fine-tuned using transfer learning techniques to improve its performance and adapt it to the specific disease prediction task. The training process involved to feeding the annotated dataset into the T5 model. The Adam optimizer with a learning rate of 0.0003 was utilized for training. The T5 model was trained for 4 epochs, and the loss was monitored on both the training and validation sets to prevent overfitting. The training procedure aimed to optimize the model’s weights and biases based on the labeled data, enabling it to learn to detect accurately.

4 Conclusion

This paper demonstrates the potential of the T5 model for automating disease prediction from unstructured medical reports. The fine-tuned model achieved high accuracy in identifying diseases, reducing diagnostic time for medical professionals. Future work will focus on expanding the dataset, improving the model’s interpretability, and exploring its integration with clinical decision-making systems to further enhance its utility in healthcare.

References

1. Kaur S, et al. Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles

and perspectives. IEEE Access. 2020;8:228049–228069.

2. Leaman R, Doˇgan RI, Lu Z. DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics. 2013;29:2909–2917.
3. Armstrong N, Hilton P. Doing diagnosis: Whether and how clinicians use a diagnostic tool of uncertain clinicalutility. Soc. Sci. Med. 2014;120:208–214.
4. Ball SA, Jaffe AJ, Crouse-Artus MS, Rounsaville BJ, O'Malley SS. Multidimensional subtypes and treatment outcome in first-time DWI offenders. Addict. Behav. 2000;25:167–181.
5. Yang Z, et al. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. Sci. Rep. 2018;8:1–9.
6. Meesala A, Paul J. Service quality, consumer satisfaction and loyalty in hospitals: Thinking for the future. J. Retail. Consum. Serv. 2018;40:261–269.
7. Shah AM, Yan X, Shah SAA, Mamirkulova G. Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. J. Ambient Intell. Humaniz. Comput. 2020;11:2925–2942.
8. Danielson B, et al. Development of indicators of the quality of radiotherapy for localized prostate cancer. Radiother. Oncol. 2011;99:29–36
9. Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J. Supercomput. 2021;77:5198–5219.
10. Koppu S, Maddikunta PKR, Srivastava G. Deep learning disease prediction model for use with intelligent robots. Comput. Electr. Eng. 2020;87:106765.
11. Noori B. Classification of customer reviews using machine learning algorithms. Appl. Artif. Intell. 2021;35:567–588.
12. Pruning, N. Measures, I. Network Pruning and Information-Entropy Measures. 1–20 (2022). [13] Radhika R, Thomas George S. Heart disease classification using machine learning techniques. J. Phys. Conf. Ser. 2021;1:012047.
13. Haraty RA, Dimishkieh M, Masud M. An enhanced k- means clustering algorithm for pattern discovery in healthcare data. Int. J. Distrib. Sens. Netw. 2015;11(6):615740.
14. Odeyemi SO, Akinpelu MA, Abdulwahab R, Ibitoye BA, Amoo AI. Evaluation of selected software packages for structural engineering works. ABUAD J. Eng. Res. Dev. 2020;3:133–141.