**The use of natural language processing in creating beneficial chatbots to help both patients and the health care system.**

The application of deep learning in chatbot in health informatics.

1. **An introduction to the problem [10 marks]**

The use of chatbots is becoming increasingly prevalent in health informatics as they prove to be beneficial in a number of ways. They are able to reduce costs, improve care patients receive and allow for patients to have an access point at times which would otherwise be inconvenient or inaccessible to contact a healthcare provider.

Deep learning is ‘subset of machine learning that uses artificial neural networks to process and analyze information’ (Google, 2024). The algorithms used in deep learning imitate the human brain. Chatbots rely on deep learning algorithms much more than machine learning algorithms because machine learning algorithms do not ‘possess the advanced natural language understanding capabilities of deep learning models thus may have difficulty interpreting user queries accurately’ (Babu and Sekhar Babu Boddu, 2024). This is beneficial as it means that chatbots are more accurate and therefore are able to respond to queries which could be more complex.

Creating a health informatics chatbots can prove challenging as a large volume of accurate high-quality data is required. It is difficult to ensure that the chatbot is able to appropriately respond to nuanced situations in which users may find themselves in. Training chatbots on deep learning algorithms means that queries which are more complex can be answered in a way that is easily understood.

1. Literature Review [30 marks]

A common deep-learning algorithm which is used when creating chatbots, regardless of their subject area, is natural language processing (NLP). The data necessary to create the chatbot is ‘learned and processed using a neural network layered with multiple layers’ (Chopde and Agrawal, 2022). This allows for chatbots to be tailored in an effective manner to meet the requirements of the user. A key benefit of deep learning algorithms is that the pre-processing steps which data undergoes allows for chatbots to be more effective as human language is understood and mimicked (Zhang et al., 2024). Chatbots greatly benefit from deep learning algorithms in several ways. The most clear benefit is the emergence of ‘personalized diagnoses and treatment recommendations’ (Rajani and Khushi Ruparel, 2023) providing a much-needed service in a climate where it is increasingly more difficult to be in contact with healthcare providers when always necessary. This links to inadequate care patients receive and how health chatbots would be able to ‘ease staff shortages and improve patient care’ (AI and Digital Regulations Service for health and social care, 2024). Limited resources alongside a growing population creates a great problem for citizens which then has a knock-on effect to the National Health Service (NHS). A chatbot may not be the entire solution to the problems which are arising but it could weed out patients which do not require immediate attention.

There needs to be an intentional, well-thought-out choice when choosing between using a deep learning algorithm and a machine learning algorithm. The key criteria which must be considered are the availability and size of the dataset(s), the run time available and the complexity of the task. Machine learning algorithms which can be used include support vector machines (SVMs) and decision trees. SMVs are ideal when performing sentiment analysis in the chatbot interactions (Meta-guide.com, 2025) and decisions trees are ideal for more simple chatbots with predefined rules (Gettalkative.com, 2024). There are various deep learning algorithms which prove useful when creating chatbots. These include but are not limited to recurrent neural networks (RNNs), transformer models such as BERT and convolution neural networks (CNNs). Neural networks are used in chatbots as they are skilled in processing natural language. They are particularly useful when in the healthcare domain because there may be multiple reasons an issue can arise. Neural network models require less statistical training, and therefore they are able to detect intricate non-linear relationships (Tu, 1996). Creating a chatbot with multiple training algorithms at its disposal permits the chatbot to provide answers in a natural manner which mimics human language.

RNNs have multiple layers and within these layers, they process the given data and store it in their memory. It then will use this initial data as context, in turn improving its accuracy (AWS, n.d.). RNNs focus on sequential data and can obtain long-term dependencies which is essential when building a chatbot. A problem that can arise with RNNs is the vanishing gradient problem. This is where the gradients between the output and input layer get smaller, leaving the weights of the layers which are lower almost unchanged (Yash, 2021). This can cause problems as it may be hard to learn long-term dependencies during the input stage which is needed for language modelling. However, a solution to this problem is long short-term memory networks which ‘prevent long-term dependency issues. Their default is retaining information for long periods of time’ (Kannan, Subbaram and Faiyazuddin, 2023).

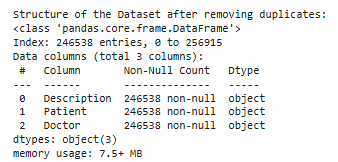
CNNs are another type of neural network which may be used by implementing a ‘word embedding’ technique, which ‘converts text into numbers in order to run text processing’ (Tsakiris et al., 2022). They are common when dealing with images, so may not be the ideal algorithm when creating a chatbot. CNNs and RNNs are similar with the prevalence of layers within the algorithm and how they will use previous layers to improve their understanding which is needed when creating a conversational chatbot.

Transformer models use a self-attention mechanism. This is when the model transforms each word in the input sequence and is associated with a query vector representing its relationship to other words, a key vector representing information about those other words, and a value vector containing the word's actual information. The mechanism computes a weighted average of the values which is determined by the similarity between the remaining two vectors. A feed-forward network is then used to generate an output by combining the weighted average and the original input. This ensures that the model can concentrate on the relevant information and capture the dependencies an RNN algorithm might have missed (h2o.ai, n.d.). Addressing this problem is highly beneficial to the clarity of the chatbot because it will provide the user with more coherent, relevant responses. Another key benefit that transformer models have is that they are able to handle running the input sequences simultaneously, therefore increasing their efficiency. A leading transformer model such as BERT provides superior results and will respond better to conversational tasks put forward by users to the chatbot.

1. **Data collection and presentation**

The health informatics chatbot was created by merging two publicly available datasets.A screenshot of a computer

Description automatically generated The dimension of first dataset ‘med.csv’ (Yousef Saeedian, 2024) is comprised of 256,916 rows of data and three columns. The relevant columns for the creation the chatbot from this dataset is the ‘Description’ column as describes the issue the patient is facing without including unnecessary information. The other useful column is the ‘Doctor’ column as this is where the answer for the given problem is given.

When investigated it was discovered that there were 10,378 duplicate rows included in the dataset which could skew any models trained on this dataset. In order to reduce any noise and redundancy within the data the duplicate rows were removed. This will improve the efficiency of the models as there will be less data to train and increase the accuracy.

A green and orange rectangles

Description automatically generated

The statistics for the word count show that the ‘description’ column is on average more brief which an average of 9.97 words in contrast with the ‘Doctor’ column with an average of 87.63 words. This makes sense as a doctor would want to be more comprehensive in their explanations to provide useful answers which are more clear.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

It can be seen that there are clear outliers within the dataset as the longest entry is 272 words for ‘Description’ and 1,867 words for ‘Doctor’. However, there is no need for these outliers to be removed as the questions or answers given could be longer depending on the issue and may just need more detail.

A close-up of words

Description automatically generatedA close-up of words

Description automatically generated

A screenshot of a computer

Description automatically generatedWord clouds for both the ‘Description’ and ‘Doctor’ variables are useful in highlighting the prominent words in the dataset. They are able to provide a quick overview and insight of the dataset.

The second dataset ‘Categories.csv’ (World Health Organization, n.d.) is made of 7 columns and 14,592 rows. The variables which are relevant to the creation of the chatbot and training of the models are ‘Synonyms’ and ‘Level 2 Health category’. These two columns are needed to categorise the descriptions from the med.csv dataset.

A screenshot of a computer

Description automatically generated

The table of category counts shows the distribution of health conditions in the med.csv dataset. There is a large number of uncategorised descriptions which may be due to an issue of missing category labels which would be relevant.

A graph of health categories

Description automatically generated

This bar chart is a useful visualisation in highlighting the key categories in the med.\_processed\_with\_categories.csv dataset which includes the med.csv dataset. The chart allows for a quick comparison and indicates what ailments are most commonly raised. This will be easily interpreted by both non-technical and technical stakeholders.

1. **Model presentation**

The deep learning model that was selected was a natural language processing (NLP) transformer, BERT-tiny. BERT-tiny was used instead of the original BERT transformer because it is able to ‘achieve competitive performances meanwhile significantly reducing the model size and inference time’ (Jiao et al., 2020) meaning it is more computationally efficient. As the dataset used had over 200,000 rows of data this was a necessary step to avoid excessive run times which would have been unachievable without extra processing power. Another choice made to aid this was choosing to use a sample of 60,000 of the dataset to optimize efficiency. The model learns by converting tokens, which are derived from the splitting of the text in the dataset, into vectors. The BERT-tiny model is bidirectional which means that it doesn’t just read text sequentially (Siva, 2021). This means that the model is able to be of more use as it will have access to more context which is useful for the chatbot understanding words in the correct context. BERT has a self-attention mechanism which means that it uses the architecture of the transformer to process input simultaneously (Dutta, 2023). Contextual understanding is key in improving the accuracy of the model. The model will also be executed faster. The weight decay the model is on is 0.1. The model learns is from the weights being updated whilst training on the dataset. This is updating the knowledge it has as the model is being trained, improving the accuracy. The model is being trained for 10 epochs meaning that the chosen sample is being passed through the model ten times in order to enhance the performance. Training with fewer epochs may lead to underfitting which would lead to less useful information given by the chatbot. It can be seen with every iteration of the epoch, both the training loss and validation loss are decreasing and gives a model accuracy of 80%.

A screenshot of a graph

Description automatically generated

Both a decision tree and a random forest algorithm were selected to analyse the collected data. For both algorithms a sample of 10,000 was used to train the models in order to be computationally efficient. The decision tree algorithm was trained with a random state of 42 to ensure results are reproducible. The max depth was set at 10 to ensure that the model must go through ten levels before making a decision, similar to the epochs in the BERT transformer. The random forest algorithm was set with 50 estimator trees. Each one of these trees learns a different subset of the dataset which reduces variance. Both of these algorithms learn by learning in each iteration of the trees, while random forest also includes feature selection. Both models were evaluated using accuracy scores and confusion matrices.

A black text on a white background

Description automatically generated



1. **Results**

The training and testing strategy of all the models was designed to ensure efficiency when processing the data, as well as showing model validation and evaluation of the performance. For the classification tasks performed by the decision tree and random forest models, variables were encoded to classify them. In all models, the data was split into 70% training and 30% testing datasets, with sample sizes of the dataset used.

The BERT transformer was amended by using the BERT-tiny variant and used the ‘trainer’ class from the Hugging Face library, with training being evaluated after every iteration of epochs. The decision tree and random forest models were trained using the TfidVectorizer technique. This is a technique where words are weighted by their importance and how often they appear.

Confusion matrices were used to evaluate the models, highlighting areas where the models did or did not perform well for the 20 top categories.

Decision tree evaluation results demonstrated satisfactory accuracy, and the classification report revealed how effectively different categories were predicted. Random Forest evaluation showed improved accuracy compared to the decision tree due to its ensemble learning approach. Both models revealed discrepancies between their predictions and the rule-based categories, providing insights into their performance.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A computer screen shot of a computer screen

Description automatically generatedThe performance of the BERT transformer model was further evaluated using AUC and ROC scores.

A graph of a curve

Description automatically generated

Results from the rule-based categorisation were compared with predictions from both the decision tree and random forest models.

A graph with red and green rectangles

Description automatically generated

When training the BERT model it can be seen that there is a clear decrease in both the validation and training loss. This highlights the model’s capacity to learn from the information it is presented with. By the seventh epoch the model begins to stabilise.

A screenshot of a graph

Description automatically generated

The BERT model had a high accuracy of 80%. This shows that the model performs well and will provide beneficial outputs in the chatbot.



The decision tree model and the random forest model were trained on a smaller sample of 10,000 due to the cost of the training times. This could mean that the decision tree model may be more likely to be subject to overfitting, whereas the random forest model is less likely to suffer from this as it benefits from ensemble learning. The decision tree model achieved an accuracy of 53% whilst the random forest model had an accuracy of 64%.

A black text on a white background

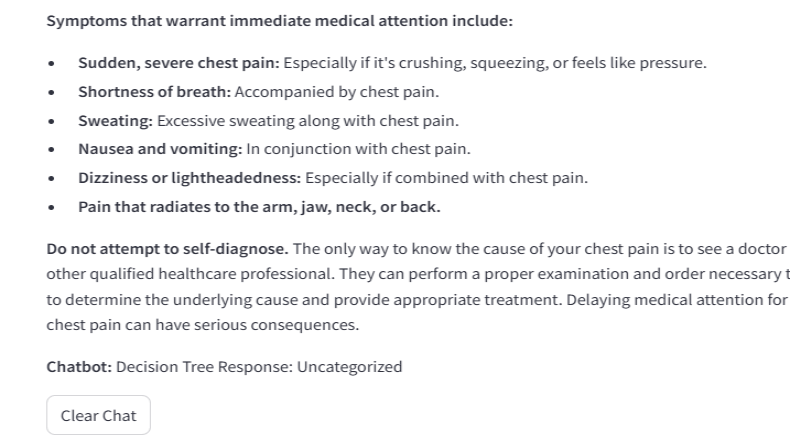
Description automatically generated

The chose model to train the chatbot is the BERT-tiny transformer as this had the highest accuracy.

Alongside this, the Google Generative AI API has been used to further enhance the BERT model. It further explain and fills in knowledge gaps that the BERT transformer isn’t able to fulfil.

A screenshot of a chatbot

Description automatically generated



**Bibliography**

AI and Digital Regulations Service for health and social care (2024). *Using an AI Chatbot to Ease Staff Shortages and Improve Patient Care*. [online] AI and Digital Regulations Service for Health and Social Care. Available at: https://www.digitalregulations.innovation.nhs.uk/case-studies/using-a-chatbot-to-ease-staff-shortages-and-improve-patient-care/ [Accessed 7 Jan. 2025].

AWS (n.d.). *What is RNN? - Recurrent Neural Networks Explained - AWS*. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/recurrent-neural-network/ [Accessed 8 Jan. 2025].

Babu, A. and Sekhar Babu Boddu (2024). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, [online] pp.100419–100419. doi:https://doi.org/10.1016/j.rcsop.2024.100419.

Chopde, A. and Agrawal, M. (2022). Chatbot Using Deep Learning. *SSRN Electronic Journal*. [online] doi:https://doi.org/10.2139/ssrn.4140506.

Dutta, M. (2023). *Unveiling the Inner Workings: a Deep Dive into BERT’s Attention Mechanism*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2023/12/berts-attention-mechanism/ [Accessed 14 Jan. 2025].

Gettalkative.com. (2024). *Decision Tree vs. AI Chatbots: What’s the difference?* [online] Available at: https://gettalkative.com/info/decision-tree-vs-ai-chatbots [Accessed 8 Jan. 2025].

Google (2024). *Deep Learning Vs Machine Learning | Google Cloud*. [online] Google Cloud. Available at: https://cloud.google.com/discover/deep-learning-vs-machine-learning?hl=en [Accessed 6 Jan. 2025].

h2o.ai. (n.d.). *What is Self-attention?* [online] Available at: https://h2o.ai/wiki/self-attention/ [Accessed 8 Jan. 2025].

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. and Liu, Q. (2020). *TinyBERT: Distilling BERT for Natural Language Understanding*. [online] Association for Computational Linguistics, pp.4163–4174. Available at: https://aclanthology.org/2020.findings-emnlp.372.pdf [Accessed 14 Jan. 2025].

Kannan, S., Subbaram, K. and Faiyazuddin, Md. (2023). *Long Short-Term Memory Network - an Overview | ScienceDirect Topics*. [online] www.sciencedirect.com. Available at: https://www.sciencedirect.com/topics/computer-science/long-short-term-memory-network [Accessed 8 Jan. 2025].

Meta-guide.com. (2025). *SVM (Support Vector Machine) & Chatbots – Meta-Guide.com*. [online] Available at: https://meta-guide.com/dialog-systems/chatbots/svm-support-vector-machine-chatbots [Accessed 8 Jan. 2025].

Rajani, G. and Khushi Ruparel (2023). Deep Learning Based Chatbot Architecture for Medical Diagnosis and Treatment Recommendation. [online] doi:https://doi.org/10.1109/icacta58201.2023.10392702.

siva, G. (2021). *BERT — Bidirectional Encoder Representations from Transformer*. [online] Medium. Available at: https://gayathri-siva.medium.com/bert-bidirectional-encoder-representations-from-transformer-8c84bd4c9021 [Accessed 14 Jan. 2025].

Tsakiris, G., Papadopoulos, C., Patrikalos, G., Kollias, K.-F., Asimopoulos, N. and Fragulis, G.F. (2022). The Development of a Chatbot Using Convolutional Neural Networks. *SHS Web of Conferences*, [online] 139, p.03009. doi:https://doi.org/10.1051/shsconf/202213903009.

Tu, J.V. (1996). Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *Journal of Clinical Epidemiology*, [online] 49(11), pp.1225–1231. doi:https://doi.org/10.1016/s0895-4356(96)00002-9.

World Health Organization (n.d.). *Health Categories*. [online] www.who.int. Available at: https://www.who.int/observatories/global-observatory-on-health-research-and-development/classifications/health-categories [Accessed 26 Dec. 2024].

Yash (2021). *The Challenge of Vanishing/Exploding Gradients in Deep Neural Networks*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/the-challenge-of-vanishing-exploding-gradients-in-deep-neural-networks/). [Accessed 7 Jan. 2025].

Yousef Saeedian (2024). *AI Medical Chatbot*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/yousefsaeedian/ai-medical-chatbot [Accessed 16 Dec. 2024].

Zhang, Y., Raymond, Jingjun David Xu, Rao, Y. and Li, Y. (2024). Business Chatbots with Deep Learning technologies: state-of-the-art, taxonomies, and Future Research Directions. *Artificial Intelligence Review*, 57(5). doi:https://doi.org/10.1007/s10462-024-10744-z.