Vinh Nguyen
300473488

AIML430 - Capstone Report

Application & Implication of Using NLP for Analysing Patient Notes

Introduction:

In the realm of healthcare, the manner in which doctors understand and interpret patient symptoms is crucial for an accurate diagnosis. Before obtaining their license, physicians have to undergo extensive training in documenting patient histories, including complaints, examination results, potential diagnoses, and follow-up care recommendations. Despite this rigorous training, the process of evaluating writing proficiency for patient notes remains dependent on feedback from peers and supervisors. Enter machine learning (ML), a tool that promises to revolutionize this age-old process.

The traditional method of having fellow physicians review and score patient notes has its drawbacks, primarily in terms of the time and resources involved. With advancements in ML technologies and algorithms, natural language processing (NLP) has emerged as the prime candidate to streamline this process. However, the unique and verified ways in which medical information is expressed present a big hurdle for computational scoring and analysis. Moreover, synthesizing information from multiple segments of text to derive medical concepts or deciphering instances of ambiguous text adds more layers of complexity to the challenge.

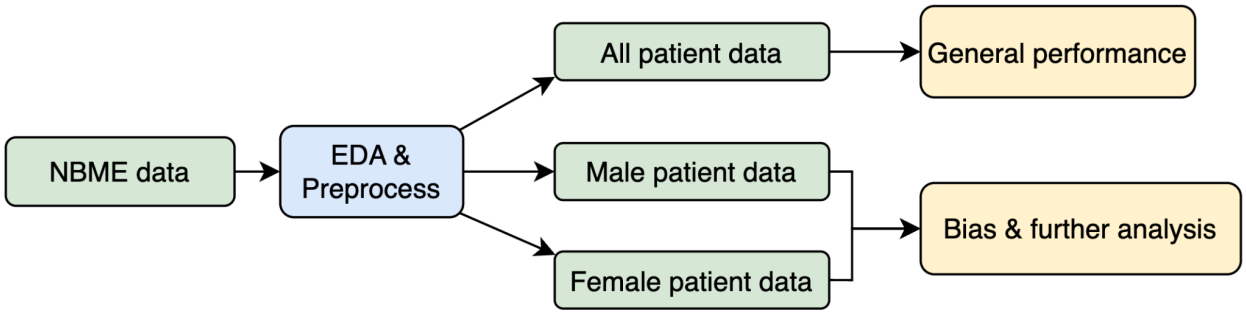
In this project, I will be investigating the effectiveness and potential biases of a state-of-the-art NLP model (DeBERTa) for automatic feature identification within each patient note. The data is acquired from a Kaggle completion **[1]** derived from the USMLE® Step 2 Clinical Skills examination **[2]**, a medical licensure exam that measures a trainee's ability to recognize important clinical facts during encounters with standardized patients. The methodology, results, discussion, and wider implications in the development and deployment of NLP in the domain of healthcare will be provided in the next sections.

Methodology:

Before running a model I first perform exploratory data analysis (EDA) to provide insights into the data characteristics (provided in EDA.ipynb). Then I apply some pre-processing steps to fix the incorrect annotation provided and split it into three different datasets: one containing the post-processed data, one containing only male patient notes, and the one containing only female patient notes.

Using Kaggle's compute server, I tested a state-of-the-art model in NLP called DeBERTa (Decoding-enhanced BERT with disentangled attention) on the three different datasets and evaluated them using cross-validation and span F1-score.

Below is an abstract view of my pipeline:



Results and Discussion:

General Performance

To evaluate the general performance of the state-of-the-art NLP model for this task, I used a baseline pre-trained model of DeBERTa. The model is evaluated by using cross-validation on the full dataset.

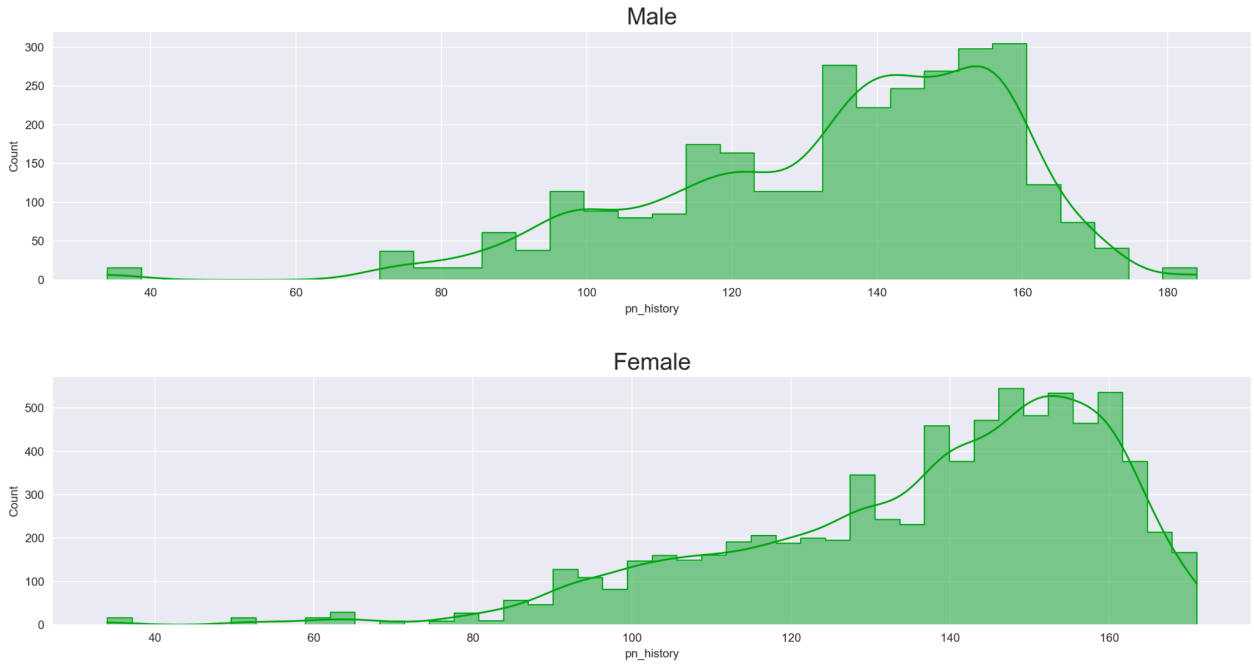
The cross-validation performance on the full data yields good performance with the full patient data. The DeBERTa achieved a good performance achieving an F1 score of around 83.7% in cross-validation. This showcases the pre-trained model's capability in comprehending and processing complex medical texts. It is important to note that the model used was only a baseline one and thus the performance could be further improved. Due to the time constraint, I was unable to test these better which would require much longer training and inference time. Nevertheless, this outcome underscores the potential of state-of-the-art NLP models in healthcare, more specifically for automatic feature extraction tasks with patient notes.

Bias investigation

To identify potential weaknesses in the model on this dataset, I split the original data into male and female patient data which then get evaluated separately. We discuss the results and provide additional analysis on this new experiment.

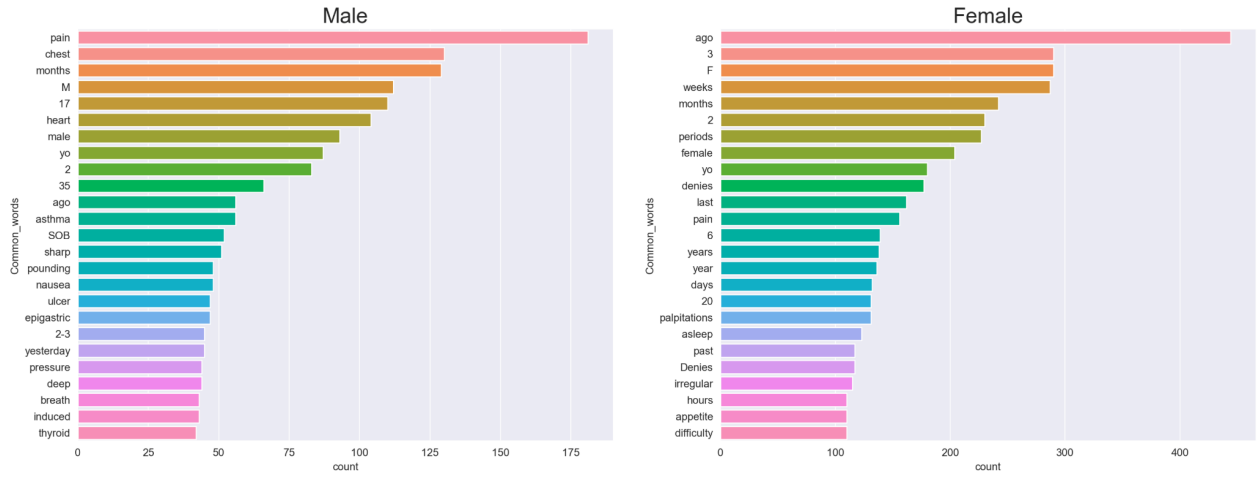
The results achieved from the model show that on average, male patient data allows for more accurate prediction than female patient data. In terms of F1 score, the model achieved around 84.2% in cross-validation for the male data and only 73.7% for the female data. At first glance, we might assume that there are biases within the model but it could also be due to the underlying characteristics of the data used. Thus, I have decided to perform additional EDA on the male and female datasets.

First, I look into the **patient notes word count distribution**:



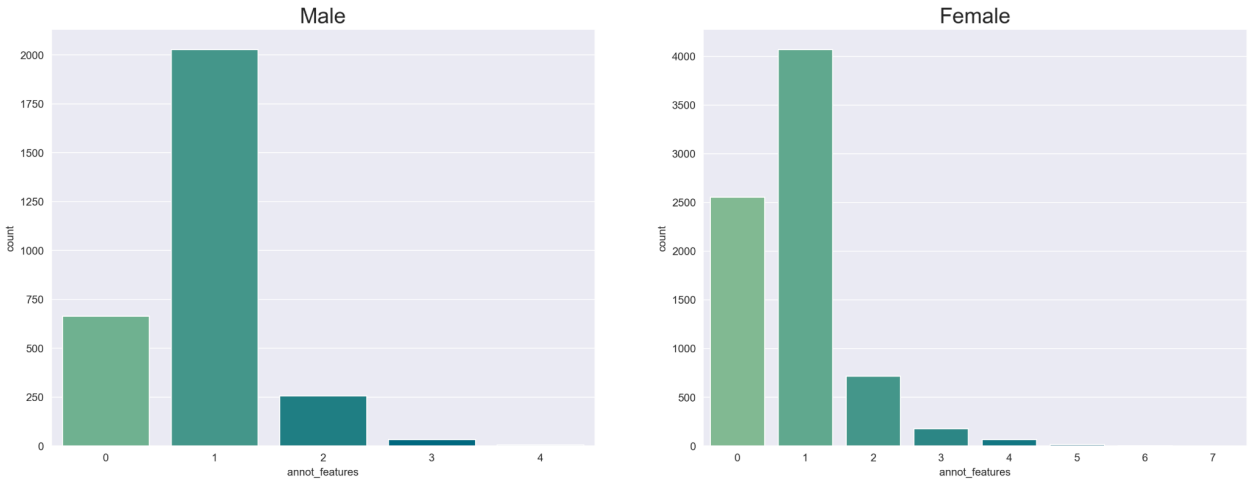
On average, the male data has fewer words compared to the female data. This might suggest that notes from male patients are more concise and information-rich, potentially enabling the model to grasp their distribution and patterns more efficiently

Second, I look into the **most common words in the annotations**:



The annotations for male data predominantly feature words related to symptoms, diseases, and specific body regions, whereas those for female data more often include common English words and numerals. Male dataset annotations appear more pertinent and distinct compared to those from female datasets. The occurrence of numerical values in the female dataset might relate to aspects like pregnancy check-ups or menstrual irregularities. Given these trends, the model is likely to more easily pinpoint key features in the notes from the male data.

Third, I look into the **distribution of number of annotations**:



Although the annotation distribution is similar for both genders, some notes from female patients may have as many as 6 to 7 annotations. This increased feature density in the notes might challenge the model, potentially leading to diminished performance.

In analyzing the dataset, I have observed indicators that suggest the presence of inherent biases. This is especially pronounced in the disparity in the model's performance between male and female datasets. The main reason for this discrepancy could be the intricate nature of the female data. The complexity of the dataset can compromise the model's ability to interpret and process it efficiently, leading to a worse performance overall. Such results can have dire consequences, especially when patient notes are involved, as any misinterpretation or oversight can directly impact patient care. Thus, ensuring a balanced and representative dataset is crucial for not only optimizing model performance but also for guaranteeing fairness in outcomes, especially in sectors as critical as healthcare.

Wider Implications:

Consent and Transparency

Patients often provide their medical data for specific purposes, such as diagnosis or treatment. Using this data for additional purposes, like training an NLP model, may not have been explicitly consented to. Ensuring that patients understand and agree to the broader uses of their data is very important. Furthermore, patients have a right to know how their data is used, including when it is processed or analyzed by an algorithm **[3]**. If patient notes are processed using NLP models to make clinical decisions, the patients should be informed about the role of these models in their care.

Bias and Fairness

If training data contain biases, NLP models can reproduce or amplify these biases (shown in the investigation above). This can lead to unequal care or misdiagnosed for certain gender or demographic groups. Ensuring that diverse patient populations are adequately represented in training data is crucial for developing an unbiased and fair algorithm or model **[4]**.

Accountability and Responsibility

Determining accountability can be difficult when a NLP model is put in place **[5]**. Is the fault with the developer, the users, the healthcare institution, or the technology itself? Clear lines of responsibility should be established before any model is put in place. This can help reinforce the safety and trustworthiness of the newly introduced technology.

Privacy and Data Security Concerns

The extraction, storage, and analysis of patient notes raise significant privacy concerns. These NLP models can be used to extract detailed information about patients for other purposes such as marketing or insurance evaluations without the patient's knowledge or consent. Ensuring the security and anonymity of this data is crucial, given the personal and sensitive nature of medical information **[6]**.

Explainability in Healthcare

The decision-making process of healthcare professionals is often based on evidence of their clinical experience. When ML-based models provide insights or recommendations without context or reasoning, there is a reluctance to trust and act on them. For clinicians to trust and integrate recommendations from an NLP model into their decision-making, they need to understand how the model arrives at its conclusions **[7]**. An opaque, “black-box” model such as NLP might face resistance in real-world adoption due to its lack of explainability and lack of trust from both doctors and patients.

Aotearoa Context

When examined in the New Zealand context, special attention to the nation’s cultural values, legal frameworks such as the Treaty of Waitangi, and social priorities is often required. While the integration of NLP models into New Zealand’s healthcare system offers exciting possibilities for improving patient

care, these models should be used in a way that protects the cultural nuances of the Maori community and enhances public trust in new technology.

Conclusion:

The use of NLP in healthcare, particularly in analyzing patient notes, can help revolutionize medical practices. By efficiently interpreting vast amounts of texture data, NLP can improve the efficiency and scalability of existing systems and provide better clinical decision-making. However, this potential is not without challenges and implications. Things such as data privacy, and ethical concerns like model bias and transparency need to be addressed before any real work can be done. Ultimately, the successful integration of NLP in healthcare depends on collaboration. Technology experts, clinicians, and ethicists need to come together to develop a holistic understanding of NLP's capabilities and limitations. Overall, NLP holds tremendous potential for improving patient care and advancing the existing healthcare system if done right.

References:

[1] N/A. (N/A). *Step 2 CK*. USMLE. https://www.usmle.org/step-exams/step-2-ck

[2] Ha Le A, Maggie, Holbrook R, Yaneva V. (2022). NBME - Score Clinical Patient Notes. Kaggle. https://kaggle.com/competitions/nbme-score-clinical-patient-notes

[3] Kiseleva A, Kotzinos D, De Hert P. (2022). Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Front Artif Intell. https://doi.org/10.3389/frai.2022.879603

[4] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. https://doi.org/10.1126/science.aax2342

[5] Habli I, Lawton T, Porter Z. (2020). Artificial intelligence in health care: accountability and safety. Bull World Health Organ. https://doi.org/10.2471/BLT.19.237487

[6] Fernández-Alemán JL, Señor IC, Lozoya PÁ, Toval A. (2013) Security and privacy in electronic health records: a systematic literature review. J Biomed Inform. https://doi.org/10.1016/j.jbi.2012.12.003

[7] Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. https://doi.org/10.1186/s12911-020-01332-6