Ailment Classification Given Description of Symptoms

Urmzd Mukhammadnaim*, Ben J. MacDonald†

FACULTY OF COMPUTER SCIENCE, DALHOUSIE UNIVERSITY, CANADA
*urmzd@dal.ca, †benjmacdonald@dal.ca

I. INTRODUCTION

A recent study has found that approximately 89% of patients across the United States research their symptoms as a preliminary action prior to seeking professional advice. Problems arise, however, when a series of articles the individual reads are contradictory. This can create a sense of confusion and concern in the user. In our research we will implement a solution to this issue by developing an algorithm that will, based on the user's text description of their symptoms, classify the most likely ailment they are experiencing.

II. NEURAL NETWORKS

Neural Networks provide a promising method of determining the user's ailment given a description of their symptoms. There are, however, several variations in which the model can be constructed. Kurup and Shetty developed a chatbot that would prompt the user to provide descriptions of their experienced symptoms. Based on this input, they utilized a sequential model with alternating dense and dropout layers followed by a softmax dense layer to classify the most likely illness. A highlighted issue, however, was in obtaining a dataset. Due to privacy concerns and a large variation in disease symptoms and severity, finding a large enough dataset to train the neural network could prove challenging. Kurup and Shetty [1] utilized a medical database which included descriptions of illnesses, however the efficacy of the model may be impacted through the use of informal terminology by users of what they are experiencing. Another model developed by Baker et al. [2] involved the use of a Convolutional Neural Network (CNN) trained with biomedical publication abstracts that would conduct multi-label classification to determine whether the text contained any of a selected ten hallmarks of cancer. In this paper, the CNN's performance was compared to that of a Support Vector Machine (SVM), and found to be more accurate. The added benefit of using a CNN would be the effective pattern recognition capabilities they provide through their convolution and pooling layers, leading to their frequent use in image and text classification. This advantage means that we are most likely going to implement a CNN as our model. There are, however, extra considerations we will have to take into account with this form of Neural Network. When constructing our model, we will have to determine the filter size used in the convolution layers. When attempting to increase the accuracy of their model, Baker et al. maintained a constant total of filters at 300 while testing the performance of their model at varying filter sizes. We will have to undergo a similar process when developing our model in order to optimize its performance.

III. N-GRAM LANGUAGE MODEL

N-gram language models have been shown to have varying degrees of effectiveness in topic categorization, authorship attribution and sentiment analysis, among other research topics [1]. In particular, an N-gram language model recently demonstrated the capability to extract general concepts from pathology reports [2]. The proof-of-concept provided by Yip et al. warrants an exploration into the use of N-gram language models in another medicine-related problem, predicting ailments from a description of

symptoms. However, the existance of a medicine-related N-gram language model does not guarantee that any meaningful results will be produced through the application of a N-gram language model in this project. This is partially due to the fact that the datasets of Yip et al. and the project's expected dataset will differ greatly in domain. To explain, the expected dataset of this project will consist of client testimonies from medical shows, which contains written text meant to resemble speech from the general public, whereas the documents used in Yip et al. are from University pathology reports, which provide domain specific information to researchers and doctors. We believe that this discrepency may prevent progress as the domain-specific language used in reports is designed to ensure that a diagnois is clearly pursued, whereas description of symptoms can vary greatly in the general public, and can often be vague without leading questions. One possible solution may involve the use of smoothing. By assigining non-zero probabilities to various n-grams, we can reduce the negative impact when applying the model in unknown contexts [3]. Further exploration will be required to determine which N-gram language model will produce the most optimal results.

IV. Next Steps

Going forward, we have developed a schedule to keep us on track for the remainder of the semester. We aim to have gathered and cleaned a usable dataset by November 12. This process may require more manual involvement in gathering usable data due to a lack of a preexisting dataset containing informal medical descriptions classified by their illness. Next, our aim is to have a completed and functioning model by November 26. One week later on December 3, we expect to have completed our report and on the following day, December 4, we want to have completed our presentation. We do plan to have started work on our report before completion of our model, however, as there will be sections that aren't dependent on information from its performance. This is to provide ourselves with enough time to create the report without a time crunch. In addition to continuous independent work, the two of us wish to call 2 times a week to bring each other up to date on our progress, as well as pair program should that be required.

References

[1] J. Kruczek, P. Kruczek, and M. Kuta, "Are n-gram categories helpful in text classification?" in *Computational Science – ICCS 2020*, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds. Cham: Springer International Publishing, 2020, pp. 524–537.

[2] V. Yip, M. Mete, U. Topaloglu, and S. Kockara, "Concept discovery for pathology reports using an n-gram model," *Summit Transl Bioinforma*, pp. 43–47, 2010.

[3] D. Jurafsky and J. H. Martin, "Speech and language processing."