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 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. involved the use of a Convolutional Neural Network 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.

III. N-GRAM LANGUAGE MODEL

N-gram language models have been shown to have varying degree of effectiveness in topic categorization, authorship attribution and sentiment analysis, among other research topics [1]. In particular, N-gram language models have 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 in this project. This is partially due to the fact that the datasets of Yip et al. and the project's expected dataset will vary in both the average word-length and the form in which they written in. To explain, the expected dataset of this project will consist of client testimonies in medical shows, which will primarly be informal, whereas the documents used

in Yip et al. are from University pathology reports, which consist of text written formally. The difference

IV. NEXT STEPS