To: John Fyfe From: Ontoadaptive Fall 2019 team Date: December 13, 2019 Subject: Final Memo \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Introduction** Elderly patients who have been diagnosed with chronic illnesses, like cancer, need to stay connected with their long distance family members in a timely, consistent and relevant manner. A patient has 5 minutes to interact with a physician during their 17 minute appointment. Family members who are geographically separated need to be informed on the patient’s health in order to best advocate for the elderly patient in the 5 minute interaction. As solution to this problem this project come to empower family. Snobird v5 is an application that empowers long distance family members to connect with patients in a timely relevant and consistent manner . Also , it helps the caregiver to communicate patient’s health condition to the physician in concise and sequential manner. The purpose of this project is to classify Whatsapp message into 5 classes that indicate the importance of the content of the message when it’s related to the health of the patient.The classification happened using deep learning model which gives the classification and produce a visualization of relevant messages on a timeline. That could be used by a caregiver to communicate the patient’s symptoms during the doctor’s visit.

**Snobird v5 Architecture**

**Input:** The messages taken from the whats app message dataset or are entered by the user through FLASK web app. **Intelligence:** The messages are then analyzed by the CNN and RNN models that provide a prediction of which class the text message belongs to (0-5).

Ontoadaptive Fall 2019 | 1

**Output**: Each message is being classified and visualized in the Plotly timeline chart with the time of the message, the classification prediction and the content of the message.

**Original Plan**

This is the gantt chart we had at the beginning of the project with the detailed tasks. This was developed after getting a prospective of what our end product would be like that happened post the discovery phase we had with our sponsor. The following are the key milestones based on the original gantt chart.

Key Milestones:

1. Creating timeline visualization for the final dashboard mockup 2. Improve current model accuracy which was developed by previous team (CNN). 3. Determining a new, more efficient deep learning model (RNN model) 4. Create plotly visualization that displays patient symptoms on a dashboard 5. Determine a platform for the Snobird App

Ontoadaptive Fall 2019 | 2

**Scope Changes**

After we began perfecting the models and we increased the accuracy through the data augmentation, our scope quickly changed after the midterm to include Google Cloud Platform deployment. As part of the agile project management simulation, our sponsor provided us with a bootcamp tracker excel sheet with three tracks we could take with the project: basic, intermediate, and advanced. We decided for ourselves which track and how many hours we wanted to put in each week. While we attempted for the advance track of the project, which is part of the reason our scope expanded, we quickly realized our lack of technical experience and some risks that the sponsor did not anticipate, for example the GCP space restrictions, this prevented us from taking the advanced track fully. In terms of the models we definitely reached the advanced track with the data augmentation and SMOTE methods, but the GCP deployment came later in the semester and we needed more time to reach that level.

Please refer to the appendix for the second gantt chart.

**Description of Final Deliverables**

**Model**

Two deep learning models were built, which are:

CNN: CNNs are generally used in computer vision, however they have shown promising results when applied to various NLP tasks as well. The model architecture of the CNN recognizes the pattern of a space .

RNN: RNNs perform very well for applications where sequential information is important because the meaning could be misinterpreted or the grammar could be incorrect if sequential information is not used. The model architecture recognizes patterns over time.

The following table represents the comparison of the two models in terms of accuracy:

Ontoadaptive Fall 2019 | 3

**CNN RNN**

Accuracy %84 %82

Accuracy with applying EDA and SMOTE

%92 %94

Because the dataset is small and imbalanced - which affects training of the model, we used two NLP techniques which is Easy Data Augmentation (EDA) and Synthetic Minority Oversampling Technique (SMOTE). This helped in addressing the issue of small dataset and imbalance thereby improving the model .

● EDA is a technique used to artificially expand the size of a training dataset by creating a modified version of the data.

● SMOTE is a statistical technique which can be used to increase the number of cases in a dataset in a balanced manner.

**Dashboard Visualization**

This is a timeline visualization which represents the classification of each message with the time and content of the message.

See this video demonstration: https://youtu.be/8ef91vop4s0

Ontoadaptive Fall 2019 | 4

**Flask App**

It is an application that works on the google platform with the model integrated within. A user can use the app to add a message and check the prediction for the same. Please see the appendix for screenshots of the flask app.

**Business Value** According to the World Health Organization, the number of people aged 65 and older is projected to grow from 524 million in 2010 to nearly 1.5 billion in 2050. Over the years, there has been an increase in life expectancy for patients as the leading cause of death shifted from infections to non communicable chronic diseases. Snobird directly targets this rapidly increasing market. It offers a solution tailored and personalized to enhance the patient journey of the aging population who have been diagnosed with chronic illness. Snobirdv5’s target market are seniors with caring families willing to pay for Snobirdv5 as a subscription.

Snobird v5 specifically has business value in the following areas:

1. Facilitates immediate access to patient data right from smartphones to all family

members. 2. Emergency cases can be handled efficiently, doctors are provided with necessary

information thus saving additional consultations and reducing costs. 3. There is an improved coordination between the physician, patient and caregiver with the

patient history data readily available.

**Next Steps**

1. BERT as a service: Google created a pre-trained model that is trained on all of Wikipedia called BERT. Bert-as-service is a sentence encoding service for mapping a variable-length sentence to a fixed-length vector, e.g. hello world to [0.1, 0.3, 0.9]. Each element of the vector should "encode" some semantics of the original sentence.

2. The dataset contains more data than just the text messages. User, date, and time data should be taken into consideration when developing Snobird v6. Example: Text messages from an important user (primary caregiver) should be flagged when the frequency is abnormal.

3. Using pre-trained embedding as a layer, example: FastText. FastText is a word embedding model where each word vector is based on sub-word characters. This means that even for a previously unseen word by the model or a typo, the model can still make an educated guess on its meaning. For this approach, the embedding needs to be handcrafted for each example since the classic Keras embedding layer will not be used.

Ontoadaptive Fall 2019 | 5

Although more time is invested into preprocessing, it is justified by better results.

4. Predict sequences, trends, symptom-based prediction - with patients like the ones with

cancer having recurrent episodes, sequences can predict what can be anticipated next.

5. When data is classified as a 3 for consecutive text messages then the classification will most likely rise to a 4 or 5. There should be a model that flags the message classified as a 3 as the start of a concerning sequence.

6. Incorporating text messaging within the Snobird app instead of the user inputting it in the

UI of the flask app.

**Conclusion**

Going into the project, none of us had experience using Python, deep learning models, or the Google platform. We were continuously challenged and expanded our technical knowledge and skill set in the process. It required a lot of outside research and guidance and resources from the sponsors. While the project was technology heavy, we also learned valuable agile management as this project was a simulation of an agile work environment. Here is our key lessons learned:

Technical:

● Learning Python

● Obtained knowledge on various concepts of Data Science such as neural networks and transfer learning

● Got a hands on learning experience on Google Cloud Platform and Google Colab

● Working with CNN and RNN models

● Techniques to work with smaller datasets - Easy Data Augmentation & SMOTE oversampling technique

Management:

● Agile project management, ex: Scrum updates, Status reports, weekly deliverables

● How to change the scope of the project over the project period

● Handling any issues that would delay the project and communicating these risks to the project sponsor

● Handling various aspects of the project and teamwork

Snobird is evolving into a potential tool that can revolutionize the way caregivers interact with physicians. The business value lies in not only empowering the family as advocates for their loved ones, but as a tool to help physicians understand patient history faster. This app as a

Ontoadaptive Fall 2019 | 6

subscription has the potential to generate exponential capital given the aging population of the United States.

**Appendix:**

**Second Gantt Chart:**

Ontoadaptive Fall 2019 | 7

**Flask App Screenshots:**

Ontoadaptive Fall 2019 | 8

Ontoadaptive Fall 2019 | 9