An Emotionally Aware Dialogue system with Memory

Vanshika Reddy
Department of Computer Science
and Software Engineering
Rose-Hulman Institute of Technology
reddyvs@rose-hulman.edu

Abstract

There has been a lot of research involving human emotions and how to interpret them. Most of this research concentrates on the identification of the response and response generation for sessions. In this paper, we propose a unique way to build a memory along with the analyzing the user's input to generate emotionally appropriate responses. Our model contains two important and distinct features, which are generating a mental model, which acts as memory for the system and analyzing the intensity of the emotion. By remembering instances and intensities of an emotional event, out model tries to etch out an emotional profile of the user which is a key input in the response generation process. The contribution of this paper is system to better understand human emotions and provide human-like emotional assistance. (put a few data pieces here to finish the abstract)

I Introduction

People are complex beings and feel a wide range of emotions. We have a need to express ourselves and, in an era where technology is becoming part of the mundane, it is important for conversational AI to respond to human emotions. Responding to human emotions can lead to increased collaboration and efficiency in HCI but for purely conversational AI, this provides the user with an assistant or a friend if one needs.

Dialogue systems such as Siri and Alexa do not do well in responding to a user's comments that carry (deep) emotions. For example, when told that "my dog has just died," the Google assistant responds with "Top ways to dispose of a dead dog's body". When told "I had a bad day," Siri responds with "You can always talk to me." However, when done so, Siri breaks the flow by saying "I'm sorry, I don't know what that means, but I can search the web for you". Clearly there is room for improvement. We are in the process of developing a system that captures and stores emotionally charged user comments and learns from them so as to react appropriately in subsequent dialogs. We note that the degree of the emotional response entirely depends on the user. We acknowledge that while one user may be in tears over the passing of their dog, another user may have a much lesser emotional response or even one of a polar opposite. Our system is designed to build a model of a user and to appropriately respond to a given user's likely emotional response.

II Background

Classifying emotions started with people trying to understand emotions and then building several frameworks to try and model their responses to emotional events. The goal of this thesis is to accomplish a level 4 emotional awareness according to the Levels of Emotional Awareness Scale (LEAS) which is comprehending blends of emotions. Though our paper focuses on generating a mental model of the user and identifying the intensity of an emotion, emotional analysis is the first and basic step.

A prominent field in which high emotional skills are required is psychiatric counselling. There has been some work in this area, with respect to creating chat bots that can behave as counselors. One such framework proposed in [7] uses a multi-modal approach consisting of several neural networks to identify emotions based on counselling sessions. This system generates responses based on high-level Natural language understanding using keywords and context. In parallel, other work used social media to derive emotions from text and generate responses [3]. These responses were not specifically for healthcare but to personalize a user's experience with the bot. A four layered Neural network was used to extract information and generate responses.

While all this work is based purely on textual analysis, another big component in understanding emotions is the recognition of para-lingual components of speech. We plan to use prior work that implements a two-branch neural network structure to analyze text along with para-lingual components to detect emotions [12]. The proposed model builds onto to it by enabling the system to associate an intensity to the identified emotion.

The Mental model

(Mental Model introduction and more work done before)

III Architecture

The basic flow of the system is as represented in Figure 1. The user sends a response, which is received by the system as audio which is then converted to text. The Emotion Recognition model processes the audio and text to identify an emotion along with an intensity rating associated with the identified emotion. The Emotion Recognition model processes and classifies the input along five indicators: Intention, emotion, context, intensity and keywords. The Intention is the direct unedited sentence the system receives from the user. The Emotion is the detected emotion from the sentence and the tone of the user. The context is a high-level summary of the emotion and keywords, which is used to check if intentions or instances repeat. The Intensity gives the system an idea of how the user feels emotionally or how the user is impacted by the situation. The keywords are indicating words from the intention that are used to build the mental model. The intention and context are directly fed into the mental model. The keywords and the context are used to check for references, previous instances and update the mental model. The emotion goes through an emotion change detector before going into the mental model to extract additional information. The mental model also uses personal information to build on itself. Personal information is used to provide background information on an instance. Personal Information includes name, birthday, family members and other important people or details.

The last two steps are Response Generation and Response Personalization. Response Generation uses the mental model, personal information and keywords to generate a response. The generated response is similar to a generic templated response. The final step consists of personalizing the response using the intention of the conversation and the keywords directly into the response before sending it back to the user.

If we take an example sentence “My dog died today” the extracted keywords are would be ‘My’, ‘dog’ and ‘died’. The Mental model would go through and check if a similar keyword exists after which a similar context is searched for. If found the Mental model knows how to respond to the user if not a new instance is made. The information such as the instance and the connected instances gets fed into response generation to create an appropriate response.

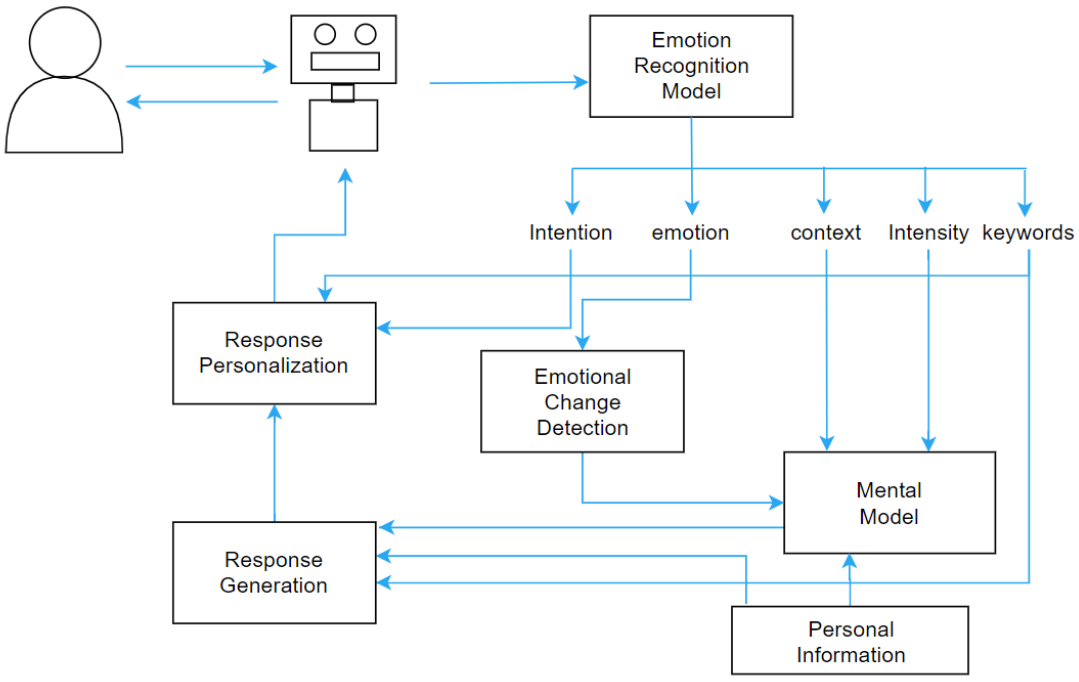

Figure 1: Proposed system architecture and flow of information.

Emotion Detection

Detecting emotions is a key aspect of our system. According to Plutchik [9], people feel a wide spectrum of emotions. The emotions can be broadly classified into 8 emotions: ecstasy, admiration, terror, amazement, grief, loathing, rage, and vigilance as shown in Figure 2.

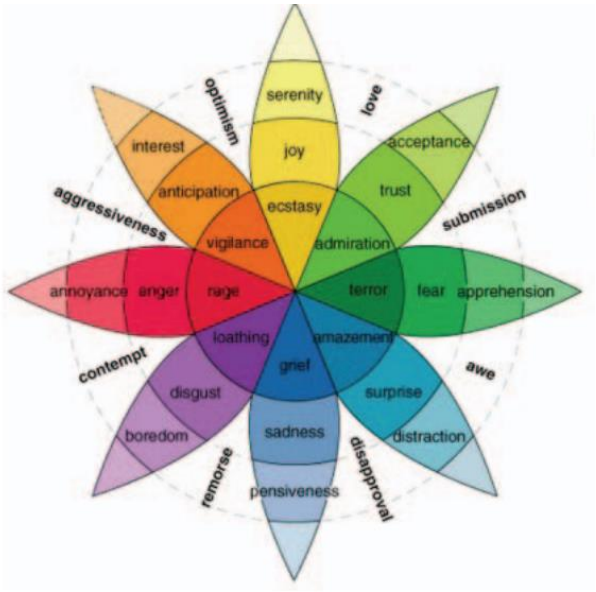

Figure 2: Plutchik Emotional Circumplex

Adapted from R. Plutchik, "Emotions and Life: Perspectives from Psychology," Biology, and Evolution, Washington, DC: American Psychological Association, 2002

We propose to use a combination of text analysis and voice frequency to detect one of the eight emotions as well as the intensity of that emotion. Each color in the picture represents a different emotion and the lighter the color gets lesser the intensity gets. In our system, when the detected emotion is grief, the system would rate the intensity of the grief on a scale from 0-10. 0 represents not at all feeling grief or neutral, 5 represents the emotion of sadness and 10, represents being despondent or grieving.

(Maybe image of CNN model)

The emotion recognition model in its current form uses a convolutional neural network and trains on a data set of 1500 audio file inputs from 24 different actors capturing 8 emotions. We included short recording of 12 male and female actors to ensure a balance in the frequency and emotions detected. The model was able to achieve an accuracy of ~70% but the focus of the model is not the emotion recognition.

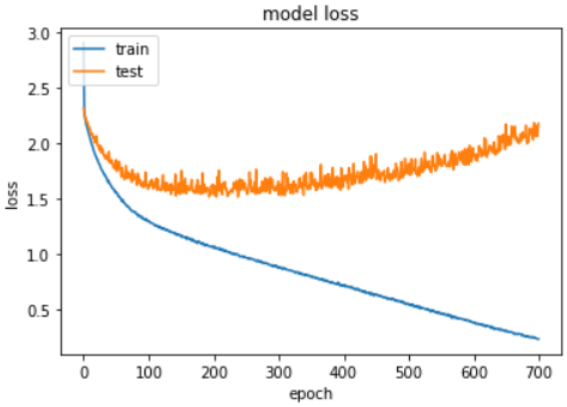

Figure 3: Results from emotion detection CNN

Intensity analyzation is also done in the emotion detection stage and due to the limitation of not having appropriately labeled data we decided to approach it using a popular unsupervised learning algorithm that is the k-means clustering algorithm. Initially when the intensity model was trained using unfiltered data and the mean and median of each data record as the features, I ended up with 10 clusters, as shown in figure 4, that didn't have a clear break down. Another issue the model encounters is that the data needs to be separated into different emotions before in order to ensure the intensity ratings match the emotion i.e. a lower frequency or pitch would mean higher intensity for the emotion sadness where as it would mean a lower intensity for the emotion happiness.

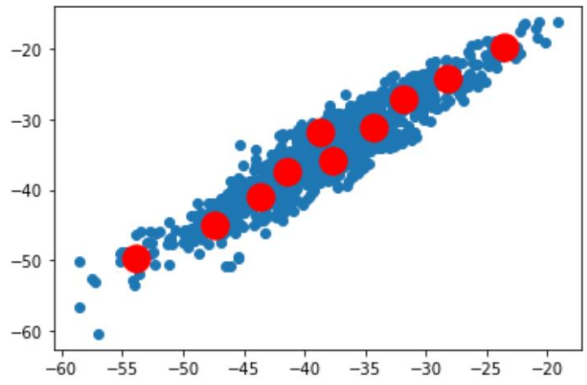

Figure 4: Initial K-Mean Clustering

The data was then cleaned using the labels i.e. data records labelled sadness were separated out just to train

the intensity model, which considerably reduces the size of the dataset to about 180 data records. This doesn't guarantee precise results and the model is more likely to suffer from the curse of dimensionality.

(need to put in final results after finding out what features get better values)

A list of emotions is defined with a number associated to each emotion in a hashed format. Every detected dialogue or instance is associated with a number that indicates what emotion it belongs to followed by a 2-digit intensity rating. The combination of the three digits gives the emotional state of the person. Since there are 10 possible intensities, 8 emotions and one neutral stage the emotional states add up to a total of 81 states. This thesis only explores 10 of the possible 81 states. The emotions and intensity rating help navigate the spectrum and to more accurately classify the emotion felt by the user. Comparing integers instead of strings is easier when looking for emotional state changes. We can also order the emotions in a way to make sure we know that for lower values we know higher intensities are bad but for higher values higher intensities are preferred. Table 1 shows the mapping along with what intensity values would be preferable for a detected emotion.

Emotion Tag	Emotion	Better intensity value
1	Greif	Lower
2	Loathing	Lower
3	Rage	Lower
4	Terror	Lower
5	Vigilance	Higher
6	Admiration	Higher
7	Amazement	Higher
8	Ecstasy	Higher

Table 1: Proposed Emotion Mapping

Using the stored emotional intensities, the model can calculate the average intensity of each emotion and model the user's usual mental stage. This work can be done and stored in the mental model which will be expanded on in further sections.

Emotional Change Detection

The Emotional change detection component is designed to recognize a user's emotions change with respect to events and the length of time a particular emotion persists. This step detects a change in the emotion and records the amount of time a user stays in that emotional state. The change in between emotions can be gradual and shift on the scale before changing emotions completely or it can be sudden. The change depends on the person as well as the event that caused the emotion and the events magnitude. It is important to remember that the emotional change detection only detects a change in the emotion and not the emotional state i.e. a difference between emotional state 110 and 108 would not be detected, but 110 and 208 would be detected as the emotion completely changes. Intensity changes will only be detected if there is a massive change in intensities or a difference of about 4 between the emotional states i.e. 110 vs 108 won't be detected but 110 vs 105 would be as |110-105| = 5 >= 4.

This unit determines the emotion, the change in intensity of that emotion and the duration of the emotion from onset to a change in emotion. The mental model records the instance of the emotional response, together with the context that caused the onset of it. This step helps provide additional insights into the user's mental model, it helps build a more accurate mental model and it helps our system with designing an appropriate response.

Mental Model

The mental model is akin to a summary of the user's mental state. The mental model is designed to capture the emotional make-up of a user. The mental model is initialized with basic information about the user. This information is extracted from an initial dialog with a given user that is part of the set-up procedure. This initial step is used to build a rudimentary mental model of a given user, a model that is continuously refined as that user interacts with the system.

In its current state, the mental model is a collection of keywords, value pairs. A new entry is generated when a keyword extracted by the Emotion Recognition Model is not currently in the data structure. To be specific, a keyword is chosen if the keyword doesn't get a close enough proximity to the existing keywords based on wordnet. For example, the vector similarity between dog and cat is above 0.8 and hence whichever keyword is seen first becomes the key and the remaining becomes a part of the value, which will be explained soon. Given a word with a significant semantic distance, such as "exam" our system produces a much lower similarity rating and would create a new entry in the data structure. Going back to our first example "My dog died today" the model would extract all 4 words as keywords. That means the model would

look through the existing structure to find an existing keyword that is similar or create an entirely new key. Assuming this is the first ever thing the user says the model would create 4 unique keys all being linked to an object that is pointing at the created instance, which is a memory that stores all the information extracted from the input along with the exact sentence given by the user, of "My dog died today".

The value associated with the keys is a linked list of instances. Each instance captures the context, exact instance, emotional state, date and time, changed emotional state, duration of emotion, a link to the instance of changed emotion, and an approved response. Each context is a few main words from the instance which encompasses the instance, or all the keywords detected from the instance. When a similar key exists, the new instance is added to the existing LinkedList of values associated with that key. An Instance reference can be under multiple keys but there is only one instance. Each individual reference has a link to the next reference pointing to a chronological sequence of instances under a key.

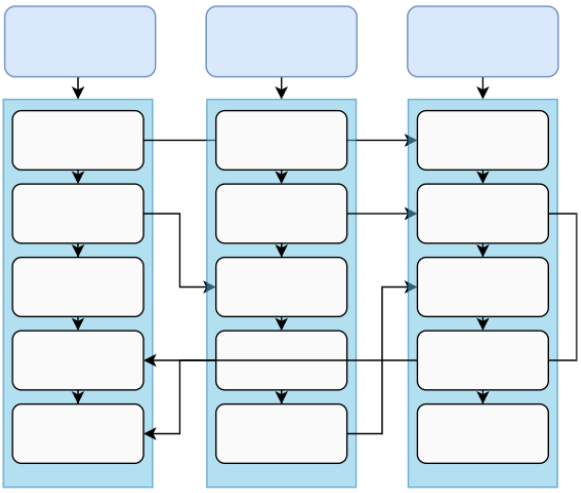

Figure 5: Mental Model internal structure

The mental model instances can be updated when an emotion change is seen or if a response becomes invalid. The mental model provides an overview of what to expect from the user's emotional profile and helps in the response generation process by providing key pieces of information and associations. The mental model is a collection of all memories that seem important and may impact how the user responds. The simplest version only catches the immediate meaning and supports the response generation process. The mental model also behaves like a data source whose data is completely self-generated and sorted.

The user's average mental state on each emotion is captured by storage variables in the mental model. Every time a new emotional state is detected the intensity is used to average the existing value of the median for each emotion. The values stored in the mental model give an understanding as to what the user's intensity in each emotion is.

By making connections to emotional state changes and associated intensities the model can find out what makes the user happy or sad or cause a change in emotion state that is in the positive direction. The mental model also happens to store the best response for situations and can use responses whenever a similar situation arises. The mental model acts as a core memory model and tries it's best to understand the user.

Another component unique to the mental model is the personal information part (shown in figure 1). The personal information stores information about the user's relations, friends, people, important dates, objects and anything that holds significant value to the user. It aids the mental model while analyzing the data received from the emotion recognition model.

Response Generation

The Response generation system is based on a preexisting chatbot provided by google. The main idea is to have a modified version of the basic available open source systems to ensure the response takes into consideration the analysis made by the mental model and the keywords. In its current state the system doesn't do that, and it would be open for future work. This step generates a generic response and ideally should use information provided to create one tailored to the situation.

Response Personalization

Response personalization is the last step before the response goes to the user. As shown in figure 1, the architecture, it uses both the intention and the keywords to make the generic response received from the generator more tailored towards the context and personalized. Though this is not one of the main focuses of the thesis there has been some work done on this. The system tries to incorporate keywords into the sentence to generate human-like responses.

IV Results

The final results of the model include accuracy calculations and functioning of the mental model. The emotional model has 2 part to account for: the emotion detector and the intensity analyzer. The CNN for emotion detector was able to achieve an accuracy of **73%** on its best run on the test set and can be improved with more data. When it comes to the intensity analyzer, it is extremely hard to figure out the accuracy of the k-means clustering algorithm. This is due to the variability in the intensity ratings. There is no defined way or absolute right intensity a person feels for a certain frequency. Some people could be more expressive while for others intensity might increase as they get quitter. The calculations on the intensity rating would be biased as they are based purely on what I think the intensity of a certain voice frequency would be.

The next part is the mental model. The main accuracy and precision calculations done on the mental model were on the keyword detection. A total of about 100 people were given 5 sentences and asked, "which words in these sentences do you think are most important in understanding the emotion of the user?". The participants rated the words as following and any word with above a 60% rating was classified as a keyword.

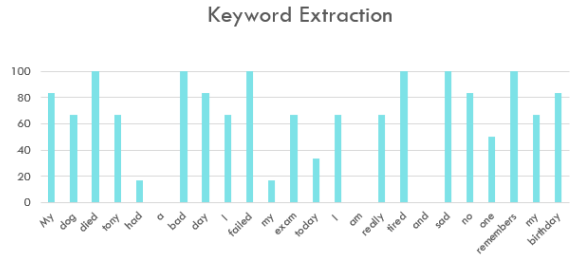

Figure 6: Keyword extraction results

That gives us the numbers for accuracy and precision as 83.3% and 87.5% respectfully. The participants were also asked to give intensity rating for each sentence and the ratings were distributed throughout the spectrum.

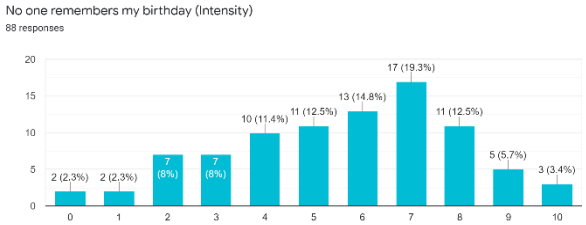

Figure 7: Intensity rating on sentence 5

As can be seen in figure 7 for the sentence "No one remembers my birthday" the intensity ratings range from 0 to 10 basically touching every intensity rating on the scale. Since the system will be tailored individually to the user it is the intensity analyzers and the mental model's job together to understand how the user would respond. Upon conducting a post questionnaire survey, most participants indicated that it was difficult to assign an intensity rating as they needed more information such as context, relation and value. That problem is addressed by the use of a personal information block.

V Conclusion and Future Work

The paper presents my senior thesis work. It outlines a model for advancing emotional awareness within dialogue systems and moves in the direction of enabling a level 4 emotional understanding. The architecture comprises of three main parts: the emotional detection model, mental model, and the response generation. The thesis focuses on the former 2 and is able to achieve an accuracy close to 73% on emotion detection and 83.3% on keyword identification.

This thesis is a step towards developing a memory and modelling the user to understand how the user reacts and responds to various emotional events. The current thesis only focuses on one emotion, but I believe the idea can be scaled to other emotions reaching a total of 81 emotional states. More work in refining the system can be done with the collection and pruning of more data to increase the accuracy of the learning algorithms.

The next immediate step would be refining intensity ratings for each individual emotion followed by working on response generation. Response generation needs to be modified to use the data provided by the mental model so it could generate appropriate responses catered towards the user.

Acknowledgements

I would like to thank and recognize Dr. Michael Wollowski of the Rose-Hulman Institute of Technology, Computer Science and Software Engineering department, for his support in the writing and editing this work

References

[1] Annabell Ho, Jeff Hancock, Adam S Miner, Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot, *Journal of Communication*, Volume 68, Issue 4, August 2018, Pages 712–733, https://doi.org/10.1093/joc/jqy026

[2] Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (n.d.). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance.

[3] Dongkeon Lee, Kyo-Joong Oh and Ho-Jin Choi, "The chatbot feels you - a counseling service using emotional response generation," 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, 2017, pp. 437-440, doi: 10.1109/BIGCOMP.2017.7881752.

[4] Endang Wahyu Pamungkas. (2019). Emotionally-Aware Chatbots: A Survey.
Ahmed Fadhil, & Gianluca Schiavo. (2019). Designing for Health Chatbots.

[5] Heleen Rutjes, Martijn C. Willemsen, and Wijnand A. IJsselsteijn. 2019. Considerations on Explainable AI and Users' Mental Models. In Where is the Human? Bridging the Gap Between AI and HCI, Workshop at CHI'19, May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, 5 pages. https://doi.org/0

[6] Johnson-Laird, P. N. (2010). Mental models and human reasoning. *PNAS*, *107*, 43rd ser., 18243-18250.

[7] K. Oh, D. Lee, B. Ko and H. Choi, "A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation," 2017 18th IEEE International Conference on Mobile Data Management (MDM), Daejeon, 2017, pp. 371-375, doi: 10.1109/MDM.2017.64.

[8] R. E. Banchs, "On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1364-1367, doi: 10.1109/APSIPA.2017.8282245.

[9] R. Plutchik, "Emotions and Life: Perspectives from Psychology," Biology, and Evolution, Washington, DC: American Psychological Association, 2002

[10] Srivastava, T. (2018). Replicating Human Memory Structures in Neural Networks to Create Precise NLU algorithms [Web log post]. Retrieved December 10, 2020, from https://www.analyticsvidhya.com/blog/2018/04/replicating-human-memory-structures-in-neural-networks-to-create-precise-nlu-algorithms/

[11] Winkler, Rainer & Söllner, Matthias: Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. 2018. - Academy of Management Annual Meeting (AOM). - Chicago, USA.

[12] Youngquist, L. M. (2020). Paralinguistic Emotional Analysis with Deep Learning (Master's thesis, Rose-Hulman Institute of Technology, 2020)

(Need to finish up the bibliography by adding all the references and making in-text citations)