Computational and Analytical Approaches to Emotion Detection

Abstract—**In this project, we explore the computational and analytical approaches to emotion detection. In the first part of the experiment, we build a Long short-term memory (LSTM) model for the task of emotion classification given the input sentences. The LSTM model achieves 81% test accuracy and has good performance on short, straightforward sentences. However, it fails to capture subtle emotions such as sarcasm. In the second part of the experiment, we have human participants determining whether the given sentences are associated with positive emotion or the negative one. There are 17 participants in total. Human participants show consistency in both responses and reasons when classifying the emotions in simple sentences, but when the emotions or structures of sentences become complex, the response and reasons vary, with most participants selecting "Cultural Background" and "Experiences" as the reasons behind their decisions, which shows the subjectivity of emotion recognition. In the discussion part, we relate emotion to cognitive science from four different perspectives: logic to emotion, rules to emotion, emotion to emotion, embodied cognition to emotion. Finally, we conclude our study by talking about the potential limitations in our experimental designs and the directions for the future study.**

Index Terms—**cognition science, natural language process, LSTM, embodied cognition, emotion**

I. Introduction

As many vocal assistants such as Alexa and Siri became popular around the world and useful to our daily life, it is necessary for these vocal assistants to detect the emotion of the speaker. With the knowledge of emotions, these vocal assistants will be able to adjust the wording, the tone, and even the content of the reply. The suitable and appropriate responses would further improve the user experiences and may further help the user in some extreme situations. Besides the vocal assistance, the online chatting platforms, social media applications, online communities, and even web applications will potentially benefit from interpreting the emotions from text. With the ability to detect users' potential emotion and expression, these applications can take actions to prevent the spread of negative emotions or the use of abusive languages. Since it is necessary to maintain a friendly environment for all users, the implementation of emotion detection would be essential.

Natural Language Processing (NLP), as an emerging area in artificial intelligence, has significant potential in processing and understanding human cognitive activities. Humans express thoughts mostly by languages. Thus, in order to understand the cognitive activities and, in particular, the emotions of humans, the focus of our natural language processing application would be analyzing the words or texts. As a hot topic that combines the psychology and the computer science field, this question has been widely discussed. Many scholars and professionals in the areas have proposed their thoughts and methods. For example, the Alm et al. [2] have utilized the sNoW learning architecture to predict the emotion from the text. Different from sNoW, another architecture has additional emotion word ontology component to assist the emotion detector in the process [3]. Besides the approaches from NLP, there are toolkit, such as EmoTxt, that could also recognize the emotion from the text [4].

Even though there are some implementations that could detect the emotion such as aggressiveness and vulgarness existed in a language, one of the limitations in these implementations is that some concepts may be neutral to one culture but offensive in the other languages. Without the cultural background, it's hard for artificial intelligence systems to decide whether certain word or expression is aggressive or not, but it is essential for machine learning algorithms to be able to grasp such emotional subtlety based on the given context. This would lead to another issue that will cause bad performance or low accuracy. Under different situation or circumstances, humans could detect different emotions even with the same sentence. However, the computational models have limited functionality for interpreting the context of sentences. Thus, the computational models would have a poor accuracy for deciding the real emotions behind sarcastic sentences or neutral sentences.

With so many information and known limitations, our research would like to address the following questions. The first question would be how will the computational model interpret human emotion. Since humans are able to detect others' emotions based on many attributes such as facial expressions, tones, body languages and so on but what attributes will the computational model focus on when analyzing emotions? Furthermore, we are also interested in the concept of society and culture within the cognitive science field. With different cultural backgrounds, people learn differently and the culture in the society will change our behavior. We decide to approach the above questions using the Natural Language Processing model to detect the emotion of the sentences. Then we will have human participants to determine the emotions contained in the same sentences.

A. Emotion

Common intuitions have it that emotions are triggered automatically, occur to people, and cause them to act in specific and diagnostic ways. For example, an offense triggers indignation. Rotten flesh triggers disgust. A barking dog triggers surprise or fear. In the midst of extreme situations, emotions come to people so rapidly that they feel that emotions overrides whatever else they might have been doing, thinking, and feeling. In scientific studies, there are many models investigating the nature of emotions. Despite the differences in their surface features, the most prominent models of emotion include the insight that emotions are automatic syndromes of behavior and bodily reactions. These models also share a common set of belief about the nature of emotion: emotions are categories with firm boundaries that can be observed in nature, either in the brain or body, and therefore recognized, not constructed, by the human mind. For example, people can sense anger just by seeing the face or listening to the voice.

With the overall landscape of theories of emotion, there are more detailed attempts and empirical records that aim to understand the underlying mechanism of emotions. Reviewed in [6], bodily activation is the idea that categories of emotion, such as anger, sadness, and fear, are distinguished by distinct patterns of autonomic response. Although some studies have reported emotion-specific patterns and behavior activation for some emotions [7], this idea is still debatable as there are also a set of studies that suggest the claim of invariance and emotion-specific bodily activity is unwarranted. The lack of emotion-related patterns that are observed in autonomic measurements can also be seen in almost all measurement modalities [6]. There is an on-going debate about whether perception-based studies of the face and voice give evidence of discrete emotion categories [9]. The variation of results in those studies is due to the fact that the experience of emotions is subjective. Despite early factor analytic evidence that self-reports produced discrete groupings of subjective experience [10], there is little consistent evidence that people routinely distinguish between feelings of anger, fear, sadness, and so on. Idiographic studies of emotion experience demonstrate that there is considerable individual variation in emotional granularity [11].

Even as scientific studies of emotion do not provide clear evidence for the biological or behavioral distinction between emotion categories, they do give clear and consistent evidence for a distinction between positive and negative affective states. Objective measurements used in the study of emotion, such as peripheral nervous system activation [12] and neural activations [13], give evidence of the intensity of a person's affective state.

B. Natural Language Processing

Emotion recognition in text is an important natural language processing (NLP) task whose solutions can be beneficial to applications in different fields, including data mining, information filtering system, and human-computer interaction. Emotion recognition in text, especially implicit emotion recognition, is one of the more difficult tasks in NLP, and it requires natural language understanding. There are different levels of text emotion recognition: document level, paragraph level, sentence level, and word level. As the level increases, the complexity of the problem increases. Furthermore, there are more subtle emotions, such as metaphors, sarcasm and irony, that are not expressed clearly in the sentences.

Different approaches have to used to recognize emotions in text. Keyword-based approaches for explicit emotion recognition have been investigated [14]. Other approaches, namely rule-based approaches [15], classical learning-based approaches [16], deep learning approaches [17]. Deep learning approaches were specifically introduced for recognizing implicit emotions in text. In our experiment, we apply the deep learning approaches to build our computational model.

II. Experiment Design

Our experiment process could be divided into two parts where the first part is implementing the computational model and the second part is designing the survey for our human participants. After both parts have completed, we will discussed the results and draw conclusions based on the data.

A. Part 1: Computational Model

The data we use for our experimentation is the sentiment140 dataset at Kaggle. It contains 1, 600, 000 tweets extracted using the twitter API. The data is stored as a csv file, which contains 6 columns (target, ids, date, flag, user, text). Relevant columns to our implementation are text and target, which means the twitter texts and associated sentiment scores labelled by annotators. The target has two potential values, 0 and 4, where 0 indicates negative emotion and 4 indicates positive emotion. Since this is an effectively binary classification, we change 4 to 1, which is a standard value used for indicating the positive category in binary classification. Since the original dataset is too large to be trained efficiently, we sample 300, 000 samples from the 1, 600, 000 tweets used for training.

During the text-preprocessing stage, we tokenize the training sentences and build a Vocabulary class that include all tokens in the training samples and in the GloVe 300 embedding, which is used for the embedding layer of our model later. Also, we include indexes for some special tokens: PADDING, BOS (Begin of Sentence), EOS (End of Sentence), UNK (Unknown).

Next, we prepare the dataset and dataloader used for training. We split the samples as 70% for training, 20% for validation, and 10% for testing. In the dataloader, we batch the data so that the training can be parallelized and therefore more efficient. Since the lengths of sentences in each batch are different, we have to pad each sentence with the PADDING token to the max length of the sentence in a batch.

With the dataset and dataloader ready, we define our training model, which is shown in figure 1. The model consists of an embedding layer, a dropout layer, a bidirectional lstm layer, and a fully-connected layer. The embedding layer is loaded with the GloVe 300 embedding mentioned previously. The dropout layer is included to reduce the impact of overfitting. The bidirectional lstm layer is used to model the context-dependent classification process, and the fully-connected layer is wired to produce the logits, which can be used to generate the final probability distribution of the positive and negative emotions.

For the training, we use the CrossEntropyLoss and Adam optimizer with the learning rate as 1e-3 for gradient descent. The number of training epochs is 7 and the batch size is 128. In order to avoid gradient explosion, we clip the gradient at 1.0. During the training, we find that the model overfits very quickly. Specifically, it overfits after the third epoch, where the training accuracy continues to increase but the validation accuracy decreases. After 7 epochs of training, the model can achieve 81.9% accuracy on the validation set and 81.3% accuracy on the test set.

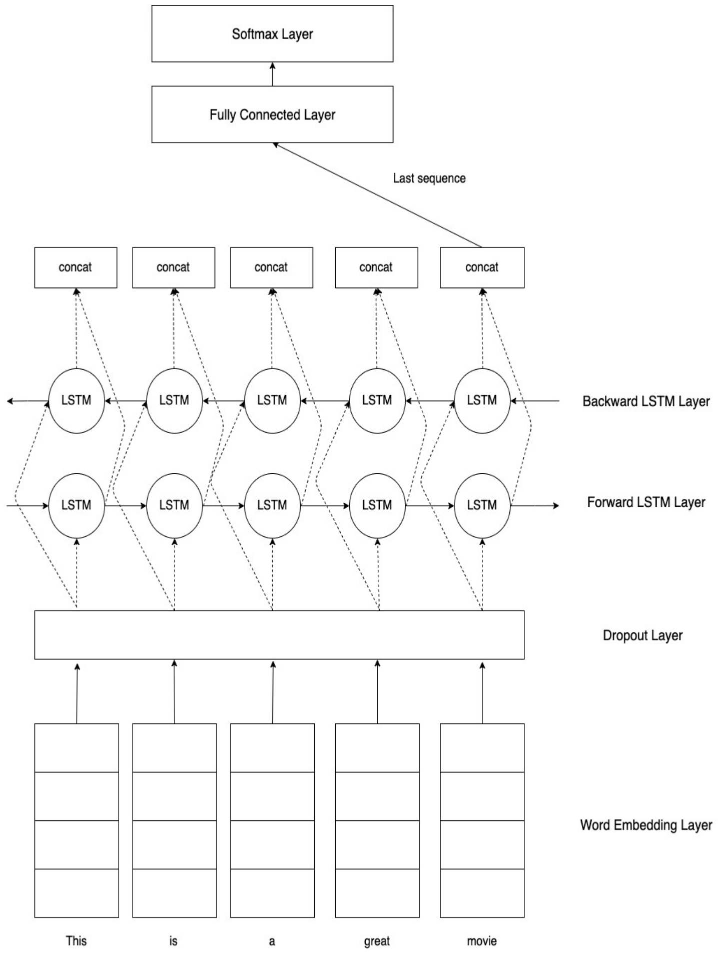


Fig. 1. Computational Model Architecture

B. Part 2: Human Participant

For the second part of our experiment, the main goal is to analyze participants' emotional reaction and reasons behind their decisions for making the classification. In order to do compare humans' performance to that of our computational model, we use the same set of sentences tested on the computational model for humans as well. After providing each sentence, we would also ask the participant to provide reasons behind their decisions. Below is the list of sentences that for both human participants and the computational model.

1) This is a great movie
2) What an awful day
3) I work 40 hours a week for us to be this poor
4) Nice perfume.
5) Nice perfume. How long did you marinate in it?
6) He wrote a meritorious theme about his visit.
7) She is friendly as a rattlesnake.

The first two sentences are designed to be the simple sentences that could filter out the invalid responses. The third sentence requires the understanding of the context from the participant. The fourth and the fifth sentence are intended for simulating sarcastic emotions. The design of putting these two sentences next to each other is to help the participants notice the sarcasm. The sixth sentence uses rare used word 'meritorious'. This is mainly to test the accuracy of the computational model and to find out whether participants know about this word. Sentence 7 is a popular sarcastic saying in English. There are many participants that we find are the international students who are proficient in English and have experiences in both English culture background and other culture.

Besides the above sentences, the computational model also tested on the following two sentences. These two sentences are relatively neutral sentences comparing with others. Even different human participants may have different opinions on the emotions contained in the two sentences, depending on their own subjectivity. When the computational model is forced to evaluate the emotion of the neutral sentences to either positive or negative, we want to find out whether the result is inclined to one side or the other.

- I am hungry.
- What is the weather today?

We distribute these questions to human participants through online survey. Based on the suggestions in [1], we keep the questions and the introductions as short as possible without creating confusion or leaving false impressions on the participants. Participants are asked to provide answers in decimal between 0 and 1 where the 0 represent the negative emotion and 1 represent the positive emotion. After answering the questions, they will be asked to explain how they make such decisions. To better classify the reasoning, instead of allowing the participants to write a short answer response, we provide possible reasons in the survey. The participants could select more than one choices that best fit to their reasons. The possible choices that we provided are:

(A) Keywords
(B) Experiences
(C) Cultural Backgrounds
(D) Intuition
(E) Other

Note the ”Other” choice where the participants could write their own reasons. This choice is included to enable the participants to precisely describe the reasons behind their decisions when none of the available choices fits.

III. Result

A. Computational Model Result

Based on the result in Fig 2, we can see that when the input sentence is short and straightforward, such as “This is a great movie”, the model can predict the correct label for the input sentence with high confidence. For instance, the model predicts with 0.91 probability that “This is a great movie” should be labeled with the positive tag. The indicative words contained in sentences, such as “great” and “awful”, play an essential role in helping the model determine whether the sentence is associated with the positive emotion or not. Nevertheless, such signature words can be misleading to the model when they are used in sarcastic sentences. For example, for ”Nice perfume. How did you marinate in it?” the model predicts that this sentence is associated with the positive emotion. Nevertheless, even though the sentence contains the word “Nice”, the sentence itself is sarcastic and thus should be predicted as negative. This exposes drawbacks of the model when doing sentiment analysis. That is, when the sentence contains subtle emotions such as sarcasm or metaphors, the model is not able to detect such emotions and make the correct prediction.

B. Participant Survey Result

The total number of participants is seventeen. According to one of the leading work about questionnaire and survey [5], it usually takes no more than 12–25 cases to reveal the major difficulties and weaknesses in a pretest questionnaire. Even though the current total number of participants for our research is not large, it should be sufficient to see the trend from the majority evaluation.

The evaluation result for the fist question are all positive, for which every participant gives a evaluation score above 0.5 and the majority of participants assign a score of 1, indicating the positive emotion of this sentence. The reasons behind this question are mainly based on the ”keyword”; about 82 percent of participants selected the ”keyword” as the reason behind their decisions and the second-most selection is the ”intuition” choice with about 53 percent participants selecting this choice. The result for the second sentence is similar to the first one. The majority of participants assigned a evaluation score below 0.5, most of which are 0. For these two questions, only a few people $(11\%, 17\%)$ select the ”cultural background” as the reason behind their decision. For the third question, about 83% participant give a low emotion evaluation score, with mostly 0 and a few 0.2, 0.3.

The results for the first three sentences are still under expectation. The results of the rest four sentences are totally different.

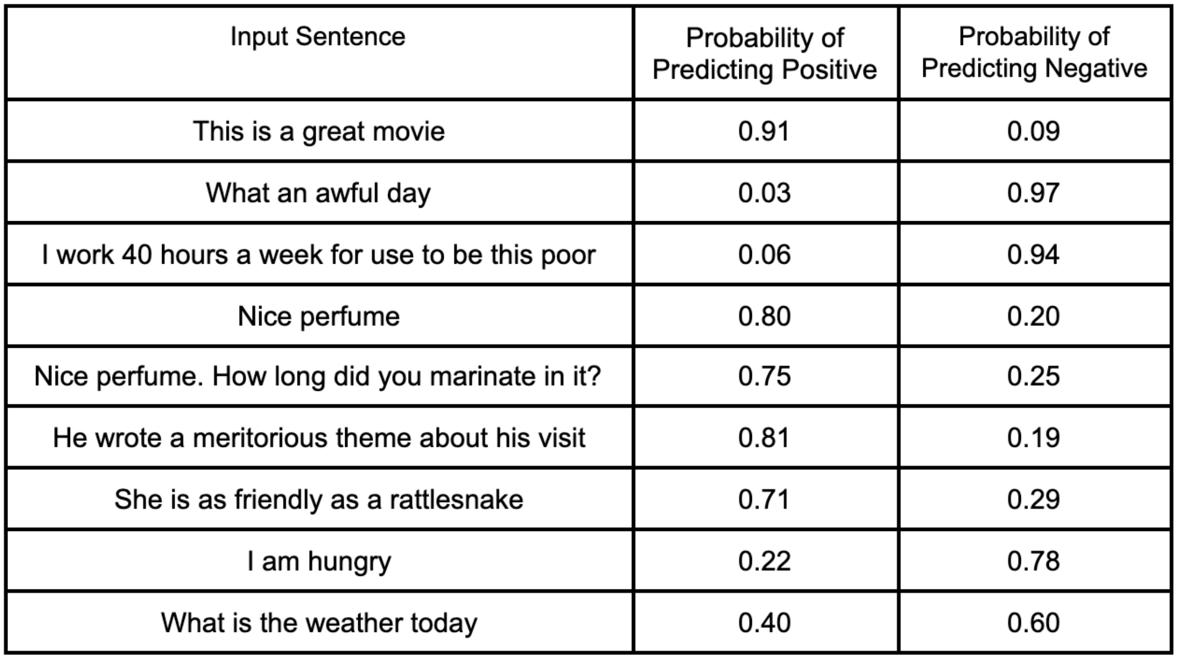
The sentence 4 and sentence 5 are a pair of sentences for the purpose of comparison. The sentence 4 is a straightforward sentence for which all participants give scores greater or equal to 0.5. The top reasons for participants are the ”Experiences” and ”Intuition.” The evaluation score for sentence 5 is the most controversial one and is widely distributed across all range. As shown in Fig.4, the responses very from 0 to 1. About 53% of participants think this sentence expresses a negative emotion where there are 35% of participants think this sentence has a associated positive emotion. This is the same situation for the response of the reasons. Every choice has more than 35% participants selected. The top 2 selections are the ”Intuition” and ”Cultural Background.” This result implies that participants think the wording of this sentence and the sarcasm behind are open to interpretations. One of the participants even gives one special reason in the ”Other” field which is ”Counter-intuitive”.

As stated in the procedure, the sixth sentence is aiming to test whether participants understand this rare used word ”meritorious”. Most participants give a rate above 0.5 and two responses detect the emotion as neutral. The main reasons for the detection of this sentence is ”Keywords”. The rest of the choices have only a few selections which roughly 4 participants.

The last sentence is designed to be an explicit sarcasm even without any context. Compared to sentence 5, the result for the last sentence is relatively straightforward. 88% of participants give a negative score, in which 41% of participants assign a 0, indicating a pure negative emotion. The reason for such selection is also clear. The leading reason is ”Keywords” with about 77% of selection. This is understandable since people possibly consider the ”rattlesnake” as the keyword that indicates negative emotion. However, the word rattlesnake by itself should not be positive or negative because it is just a noun that represents a kind of snake. This shows that even without the context for the sentence, humans are still able to detect the usage of the words and detect different emotions even for the same word.

C. Integrated Analysis

For people with different perspectives, they may also identify different emotions for the same context. For example, a person who hates Tom may gave neutral to positive emotion to a context taunting Tom, but a person who loves Tom would do the otherwise. Different from NLP models, humans internally bring various biases and presumptions when they process context. A well-trained reader is able to minimize the bias and take the neural position. However, this special case should not be applied to everyone. For each participant, their reading abilities vary. Advanced and well-trained participants can locate key words and identify emotions accurately in a short time. However, it’s relatively hard to measure this kind of bias. The ways to determine how much bias a participant has and whether the experiment should study participants with bias are questions we need to further investigate.



Input Sentence	Probability of Predicting Positive	Probability of Predicting Negative
This is a great movie	0.91	0.09
What an awful day	0.03	0.97
I work 40 hours a week for use to be this poor	0.06	0.94
Nice perfume	0.80	0.20
Nice perfume. How long did you marinate in it?	0.75	0.25
He wrote a meritorious theme about his visit	0.81	0.19
She is as friendly as a rattlesnake	0.71	0.29
I am hungry	0.22	0.78
What is the weather today	0.40	0.60

Fig. 2. Predictions by the Trained Model on Example Sentences

IV. Discussion

A. Logic to emotion

People make decisions partly based on their logical judgement. When they read a passage, they tend to make a conclusion or find out the main point by first locating some key words and key sentences. The similar situation occurs when the participants in the experiment try to decide emotion in written words. For example, the participants tend to have the logic pattern: if A, then B. When participants see some negative words, they tend to believe the passage has negative emotion. If two negative words and one positive word appear in the same passage, the participants generally try to find out the internal logic of the passage.

Logic can be extended beyond the basic logic rules. When people read a passage, they need to grasp the logic flow if they want to truly comprehend the author's idea. For example, a qualified reader should find out the main point, the supporting materials, the sub-points, and how those elements connect with each other. Our research is about emotion. The similar study can be how people identify the author's attitude. To identify the attitude, people need to have a comprehensive understanding regarding the passage. To have a comprehensive understanding requires people to think in logical ways.

B. Rules to emotion

In the experiment, participants who understand the emotions of passages correctly apply rules. They first need to learn and memorize specific language patterns, words, and expressions. Participants need to understand them, store them into memory, and know how to apply them in other different cases. From the cultural perspective, emotions are can be felt regardless of the readers' cultural backgrounds. But that does not counter our argument. For example, most people will not have positive emotions when they see the word wolf. How do they build the connection between wolf and non-positive emotion? Do they all see the alive wolf or have they been attacked by the wolf? Probably not. People learn this connection by reading books about how dangerous the wolf is or by seeing a movie about wolf hunting other animals. People reinforce this connection throughout their life. This process is demonstrated in figure 3, where reasoning, leaning, and memory are interconnected and affects each other.

C. Emotion to emotion

We study how people perceive emotions in passage, which is different from studying emotion itself. For emotion itself, there are four theories explaining it: James-Lange Theory, Cannon-Bard Theory, Schachter-Singer Theory, Lazarus Theory. The first theory argues that an event causes physiological arousal first and interprets this arousal as an emotion. In our study, participants need to experience this physiological arousal in order to detect emotions of the passage.

For the second theory, it argues that physiological arousal and emotion are triggered at the same time. This situation occurs when the participants arousal emotions by the passage itself instead of conducting emotion from physiological arousal. The third theory requires to identify a reason for the physiological arousal.

Emotions usually can be perceived by visual, auditory, olfactory, and physiological sensory. For example, people perceive others' emotion by observing their facial expression. Those factors can be represented into the passage. When people read the passage, they are aroused emotion by having

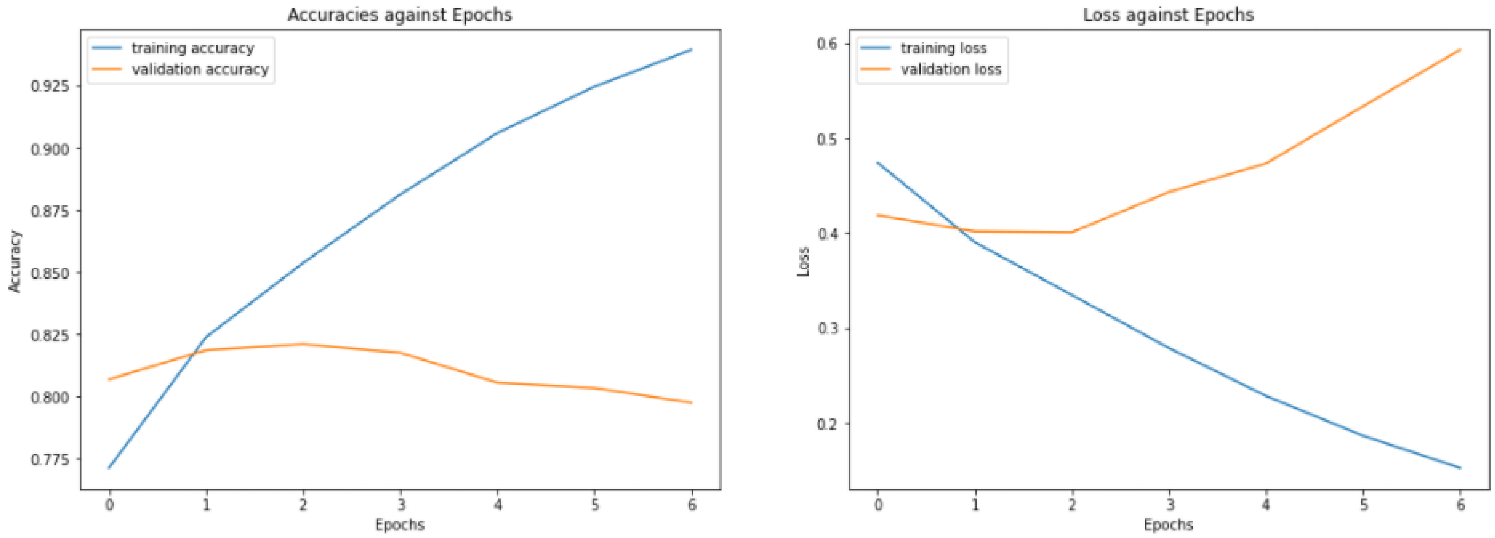


Fig. 3. Accuracy and Loss Result

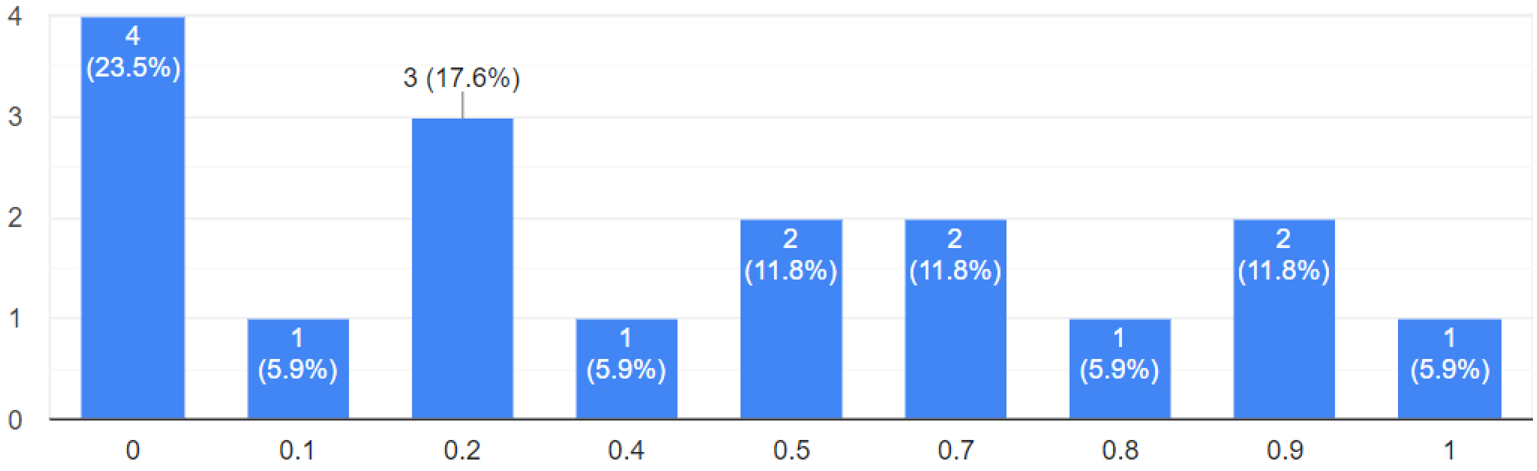


Fig. 4. Evaluation Result for Question 5

visual, auditory, olfactory, or physiological sense triggered by context.

D. Embodied cognition to emotion

Embodied cognition argues that many features of cognition are affected by the body of the organism. Bodies have the motor system and the perceptual system, interacting with the environment. The embodied cognition perspective revolves around the embodied language processing. In the theory of embodied semantics, the meaning of words is tied with the sensory motor processing unit associated with the word meaning, which is called semantic hubs. Participants in our experiments are asked to read the given texts and identify emotions contained in these texts. This process involves understanding the meaning of words and the embodied semantics. However, participants sometimes get errors. They may have different answers to the same text. There are two kinds of reasons that attempt to explain such mistakes. First, it is possible that people mistakenly process the words and connect the wrong unit in their mind. Second, it is also possible that people build different connections between words and motor units.

V. Conclusion

In this project, we use both computational and analytical methods to analyze how machine and human recognize emotinos. In the first part of the experiment, we train our computational model on the sentiment140 dataset from Kaggle. Even though our model can correctly classify the emotion in

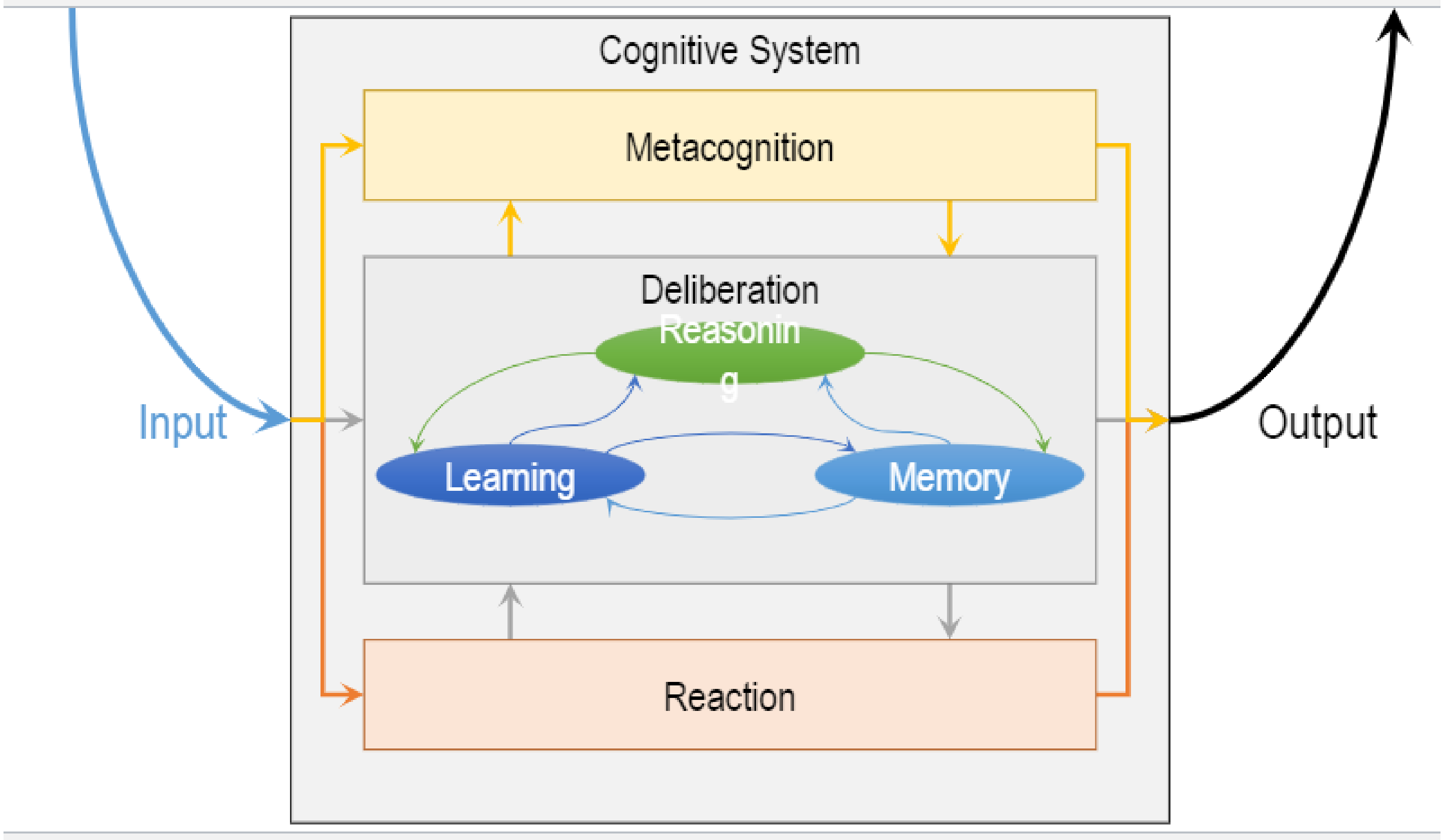


Fig. 5. Cognitive Architecture[17]

simple sentences, it fails to detect the more subtle emotions in structurally complex sentences. This shows that our model mainly uses keywords as the basis for classification. In the second part of our experiment, we give out participants the same set of sentences that are used to test the model and ask them to identify the emotion contained in each question. For simple sentences, people tends to use keyword recognition as the basis for emotion recognition. For complex sentences or sentences with subtle emotions, people tend to rely on their personal experience and cultural background for emotion classification. Therefore, the reasons behind their decisions vary, and the classification is more subjective.

In the experiment that involves experiment, we also find that emotion recognition made by humans differ from each other when they deal with some texts. To explain such phenomenon, we discuss the the process of emotion recognition from several perspectives. The first is logic, where we argue that people need logic to decide emotion. At least in some cases, logic is required for people to detect the emotion in text. The second is the rules, in which we argue that people need rules when they decide emotions in texts. They would build up patterns and rules in their minds and use those rules to decide the emotions contained in the texts. The third perspective comes from emotion itself. People need to have emotion first and build up the connection between their own emotion and the emotion in texts. The fourth thing is embodied cognition. These four perspectives are also connected to the cognitive theories learned in the class as well. Specifically, Logic and rules are included in CRUM (Computational Representation Understanding of Mind). The other two perspectives (emotion and embodied cognition) are included as possibilities to expand CRUM for better understanding humans' emotions.

VI. Limitation

A. human

Humans have their own limitations which will eventually affect the experiment. Some limitations are inevitable. There are three facets regarding the limitation of our experiment involving human participants. There are three human limitations in general. First, humans have various presumptions before they participate in the experiment. Second, humans make inevitable mistakes when they recognize emotions. Third, human describe and report their result subjectively.

Presumptions affect people's decisions. People with different presumptions will make various biased decisions, which lead the experiment result to be confusing. Those presumptions are formed throughout participants daily life. Most participants will not be aware of those presumption in the past so that they will not inform the researchers during the experiment. For example, a participant who presumes that the goat represents the devil tends to have the negative feeling when this participant sees the word goat. Another participant might thinks that the goat is a cute and kind animal so he/she might has positive emotion toward this word.

People also make many mistakes, and it occurs randomly. It is hard to distinguish random mistakes and errors from presumption-based mistakes. During the experiment, participants might feel annoyed and make their decisions randomly. Even if every participant takes this experiment very seriously and do their best in answering questions, they also might make mistakes if they ignore some important information in the passage or they wrongly conduct their logic. For participants themselves, they can neither control this uncertainty nor notice their mistakes to researchers. In this case, researchers always find confusing experiment results which cannot be explained. For researchers, we can define those abnormal data as the noisy data and ignore them. However, if we do so, we compromise the integrity of the data.

The third point is subjectivity. Participants describe the reasons for them to make decisions in subjective ways. Their descriptions can not be scaled in any objective measures. Participants might miss some important points which lead them to a more sensible decision. However, from the other perspective, emotion is also a subjective thing. People are unnecessary to have same emotions regarding the same situation. People conduct and describe their feelings in their own way. Sometimes they may not be able to express that clearly.

Furthermore, we confine the emotion in our research to be either negative or positive. However, emotion is a complex reaction in human body. It's various and complicated. Some emotions, for example, are neither positive nor negative, and they are even hard to be described by language. In further research, if participants are asked to identify emotions in a long passage, in which the emotion is complex, the participants may not be able to provide an accurate description of the emotion they detect.

B. *Limits for Computational Model*

As mentioned in the result section that our model worked well for the common used and easily understandable sentences. From another perspective, this also implies that our model may not be able to identify the languages or wordings that are not widely used in daily life such as 'meritorious' could represent good, 'inauspicious' would be used in expressing a less positive situation. For fancy or rarely used words like these, they are not often used in daily chatting which indeed may not be included in the training data set. The lack of those words during the training process would further lead to a lower accuracy for testing.

Moreover, the way our model handles the unknown word based on the various testing indicates that the model tends to give the negative emotion a little higher possibility than the positive emotion rather than assigning a neutral possibility for both, which ideally should be fifty fifty. For example, if the given input is 'Weather' which humans would not classify as either positive or negative but the computational model hasn't seen before, our model would identify this word with a 53 percent negative emotion and 47 percent positive.

Similarly, when the emotion or the intention of the sentence could only be understood with the content such as sarcastic sentences or ironically speaking of something, unlike humans, the model would not be able to understand the real meaning behind it. Computational models will treat those sentences with irony as praise and give a positive result such as classifying 'you are friendly as a rattlesnake' as a positive sentence whereas humans could clearly see the sarcasm meaning behind.

There are two limitations for our models that we would like to improve in the future study. The first one would be the number of datasets that we could fit into our model. As mentioned in the processing, we found a dataset with roughly 1.6 million data. However, with the limitation of the computing power, we could not train our model with such large data in a reasonable amount of time. We reduce the size of the dataset to 30 thousand after several trials. With too little data, the number of words the model has been exposed to will be limited and further causes a lower accuracy for testing. So far, 30 thousands of data could provide a relatively good result and accuracy for our testing but the more data could be used the better the result would be.

Even with this number of data to train our model, we are still facing an overfitting issue. After the third epoch, the difference between validation accuracy and the training accuracy grows larger and larger for future epoches as shown in Fig. 5. This overfitting situation could not provide the generalization for our model which is not what we want. There are many reasons that may cause this. One possibility may be our limited data size or it may be due to the size of the training set being too large. The other possible reason could be there are too many network layers in our model.

C. *Future study*

Exploring people's performance in more complex sentences is the first thing we are going to do in the future. The more complex a sentence is, the harder it is to identify the associated emotion because there are more irrelevant information which will disturb people's discernment. On the other hand, in order to classify the emotion associated with a complex sentence, we would need a more capable NLP model, which we will also work on in the future.

We want to decrease the impact of errors, biases, and presumptions made by human. Before the experiment, we will try to find participants with better reading capabilities and ask them if they can pay more attention on answering questions. We would find a comfortable and closed experiment room and give instructions to participants if necessary. In such a experiment environment, we can diminish the influence from other irrelevant factors and get a experiment result which reflects the real situation.

For the survey, we would reach out to more participants with diverse backgrounds so that we can gather representative responses. Then we will separate the response between the participants with different cultural backgrounds. Currently, the responses for two groups of participants (multiple cultural backgrounds and single cultural backgrounds) are mixed up together. The first approach to solve this problem is to add one more question that asks for the participants' cultural experiences such that participants will self-identify themselves. The other option would be designing different kinds of survey for different groups. For each version of survey, the questions will be altered to suit for the specific culture.

Finally, one of the issue that we find in the survey is that the ordering of the question does have certain effects on the participants' evaluation. We would also like to find out how the ordering influences the evaluation scores of the emotion and the reasons behind. To this end, we expect to have multiple version of survey with different ordering of the questions to test out the influences brought by the ordering.

References

[1] J. Converse, S.Presser, "Survey Questions: Handcrafting the Standardized Questionnaire," Quantitative Applications in Social Science, series 63, 1986.

[2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," In Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp. 579-586. 2005.

[3] S.N. Shivhare, S. Khethawat, "Emotion detection from text," arXiv preprint arXiv:1205.4944., 2012 May 22.

[4] F. Calefato, F. Lanubile and N. Novielli, "EmoTxt: A toolkit for emotion recognition from text," 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 79-80, 2017, doi: 10.1109/ACIIW.2017.8272591.

[5] P. Sheatsley, "Questionnaire Construction and Item Writing," In Handbook of Survey Research, ed. Peter Rossi, James Wright, and Andy Anderson, pp. 195–230. New York: Academic Press, 1983.

[6] Barrett, Lisa. (2006). Are Emotions Natural Kinds?. Perspectives on Psychological Science. 1. 28-58. 10.1111/j.1745-6916.2006.00003.x.

[7] Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The Tie That Binds? Coherence Among Emotion Experience, Behavior, and Physiology. Emotion, 5(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

[8] Levenson, R. W., Ekman, P., Heider, K., & Friesen, W. V. (1992). Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. Journal of Personality and Social Psychology, 62(6), 972–988. https://doi.org/10.1037/0022-3514.62.6.972

[9] Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. Psychological Bulletin, 115(2), 268–287. https://doi.org/10.1037/0033-2909.115.2.268

[10] Nowlis, V. (1965). Research with the Mood Adjective Check List. In S. S. Tompkins & C. E. Izard (Eds.), Affect, cognition, and personality: Empirical studies. Springer.

[11] Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. Journal of Personality and Social Psychology, 69(1), 153–166. https://doi.org/10.1037/0022-3514.69.1.153

[12] Bradley, M. M., & Lang, P. J. (2000). Measuring emotion: Behavior, feeling, and physiology. In R. D. Lane & L. Nadel (Eds.), Series in affective science. Cognitive neuroscience of emotion (p. 242–276). Oxford University Press.

[13] Barrett, L. F., & Wager, T. D. (2006). The Structure of Emotion: Evidence From Neuroimaging Studies. Current Directions in Psychological Science, 15(2), 79–83. https://doi.org/10.1111/j.0963-7214.2006.00411.x

[14] Ma C, Prendinger H, Ishizuka M (2005) Emotion estimation and reasoning based on affective textual interaction. In: Tao J, Tieniu T, Picard RW (eds) Affective computing and intelligent interaction. Springer, Berlin, pp 622–628

[15] Lee SYM, Chen Y, Huang CR (2010) A text-driven rule-based system for emotion cause detection. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, CAAGET'10. Association for Computational Linguistics, Stroudsburg, pp 45–53

[16] Alm CO, Roth D, Sproat R (2005) Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, HLT '05, pp 579–586Return to ref 4 in article

[17] Baziotis C, Nikolaos A, Chronopoulou A, Kolovou A, Paraskevopoulos G, Ellinas N, Narayanan S, Potamianos A (2018) NTUA-SLP at SemEval-2018 task 1: predicting affective content in tweets with deep attentive rnns and transfer learning. In: Proceedings of The 12th international workshop on semantic evaluation. Association for Computational Linguistics, pp 245–255

[18] Sweller, J., Van Merrienboer, J. J., Paas, F. G. (1998). Cognitive architecture and instructional design. Educational psychology review, 10(3), 251-296.