Beyond Modeling Of Categorical Emotions In A Neural Network Based Social Chatbot

Master Thesis Colloquium

26 February 2019

Author:

Harshita Jhavar

Examiners:

Prof. Dr. Dietrich Klakow,

Prof. Dr. Vera Demberg

UNIVERSITÄT DES SAARLANDES

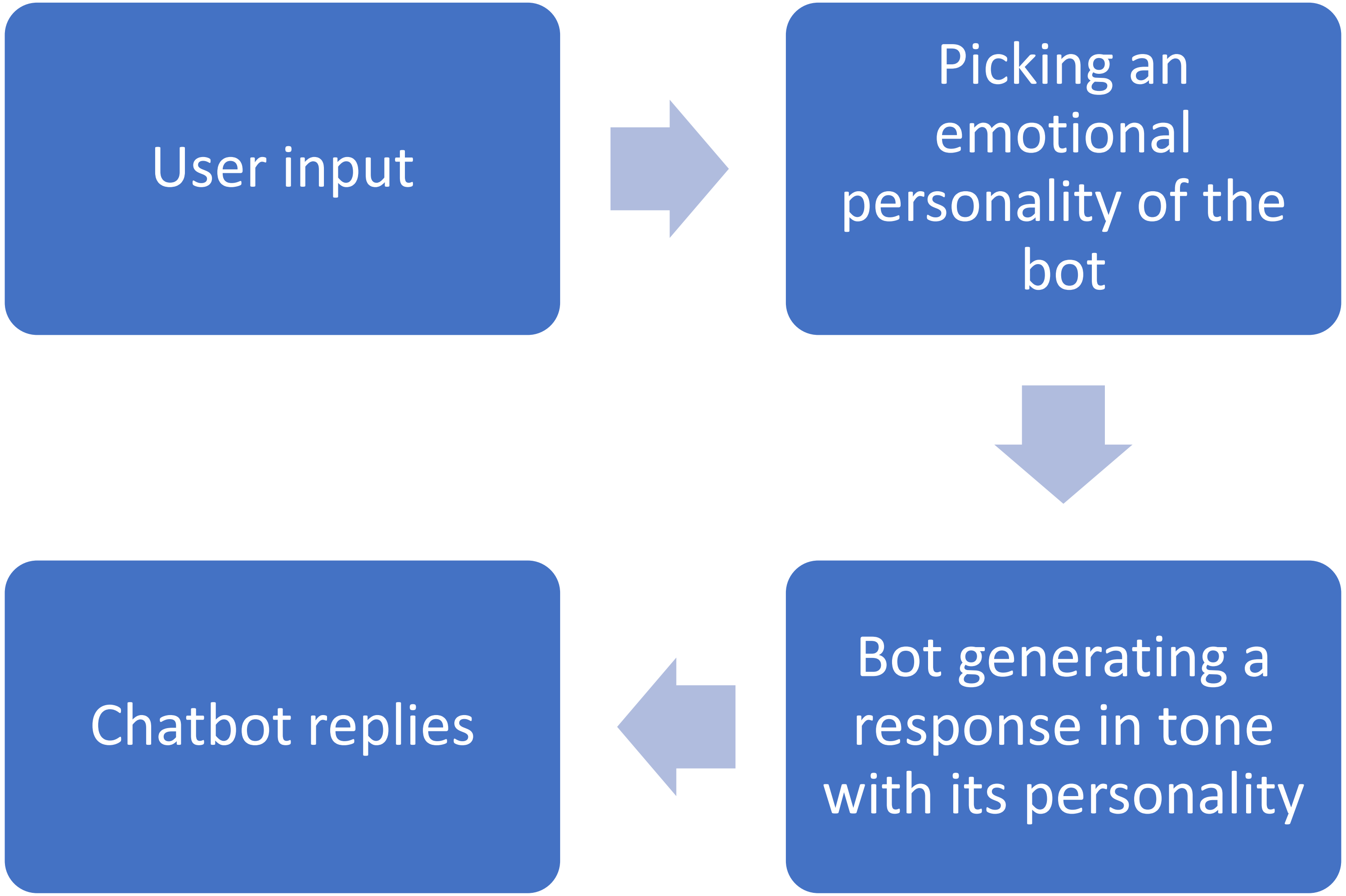
Overview

- Motivation
- Steps Involved in Emotional Response generation: 3 Goals
- Emotion Theory
- Literature Survey
- Dataset
- Task 1: Developing a mapping between categorical and dimensional model of emotions
- Task 2: Emotion analysis of the user utterance
- Task 3: Generate emotional responses from the chatbot
- Conclusion
- Future Work

Motivation

- Develop technologies based on human-like conversational models
- Applications that are inline to human's multi-sensory capabilities
- We use the same set of words to express different quality of emotions. Every sentence shares a percentage of emotions shared.
- Generating emotional responses improves the overall user experience
- Incorporating categorical and dimensional set of emotions to emotional conversational model.

Steps Involved In Emotional Response Generation



Steps Involved In Emotional Response Generation

User input

Picking an emotional personality of the bot

Goal 1

Building a new categorical + dimensional emotions scored word corpus

Emotion Analysis

Goal 2

Bot generating a response according to chosen emotion passed in NN Architecture

Goal 3

Feeding Goal 2 corpus in a NN architecture to generate responses

Chatbot replies

Emotion Theory

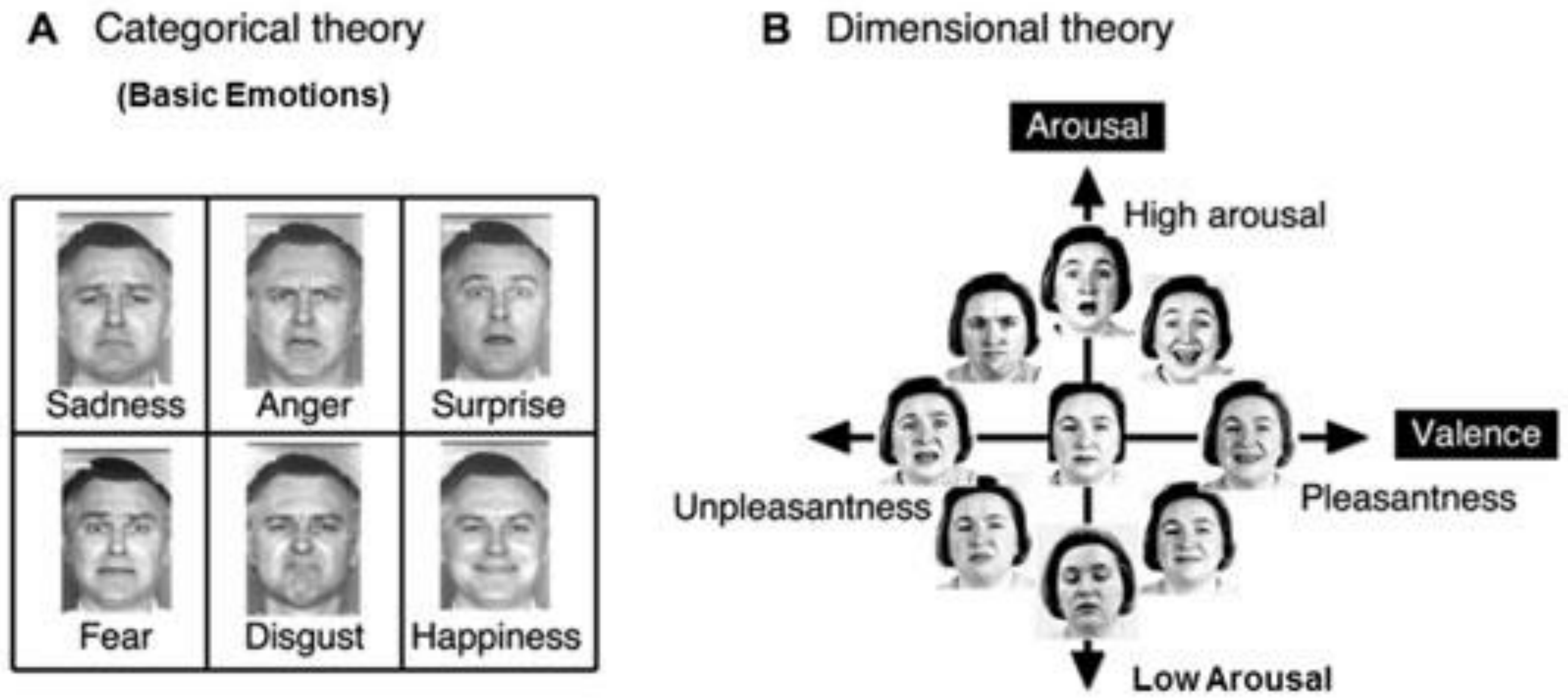


Image Source: Matsuda et al. "The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception." Frontiers in human neuroscience 7 (2013).

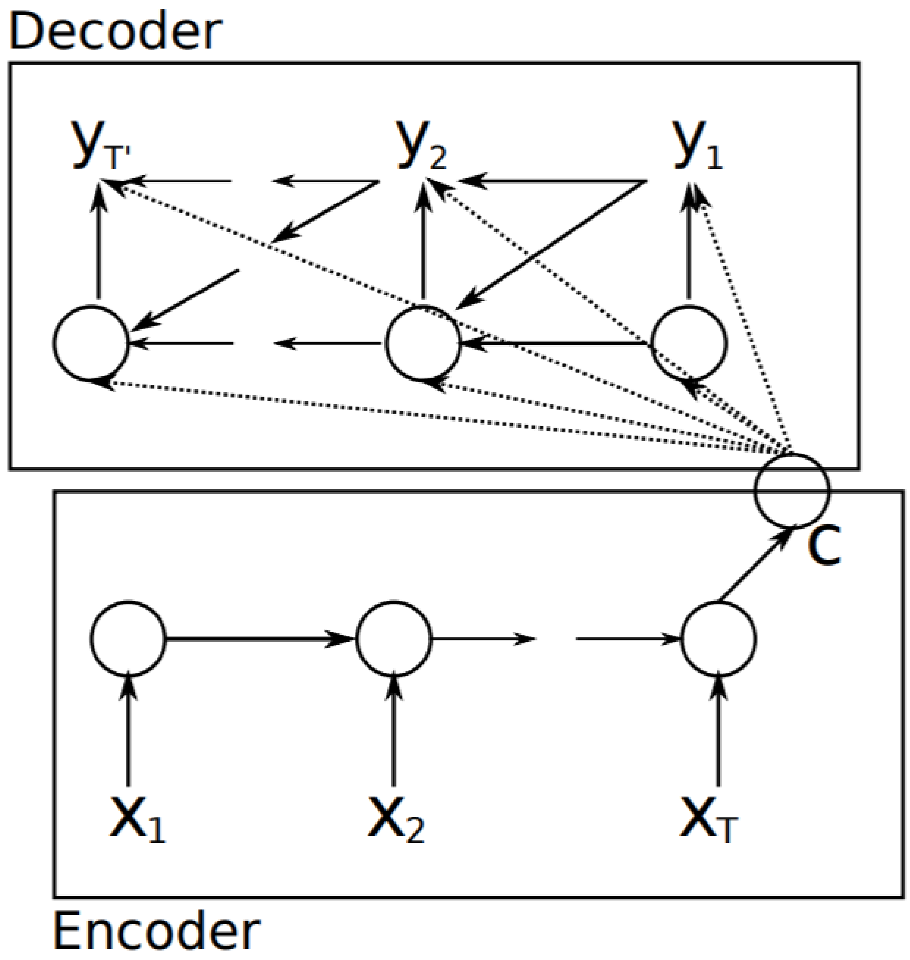
Literature Survey

Source: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation **[s2s]**

Kyunghyun et al. 2014

Theano (official), Pytorch, Tensorflow

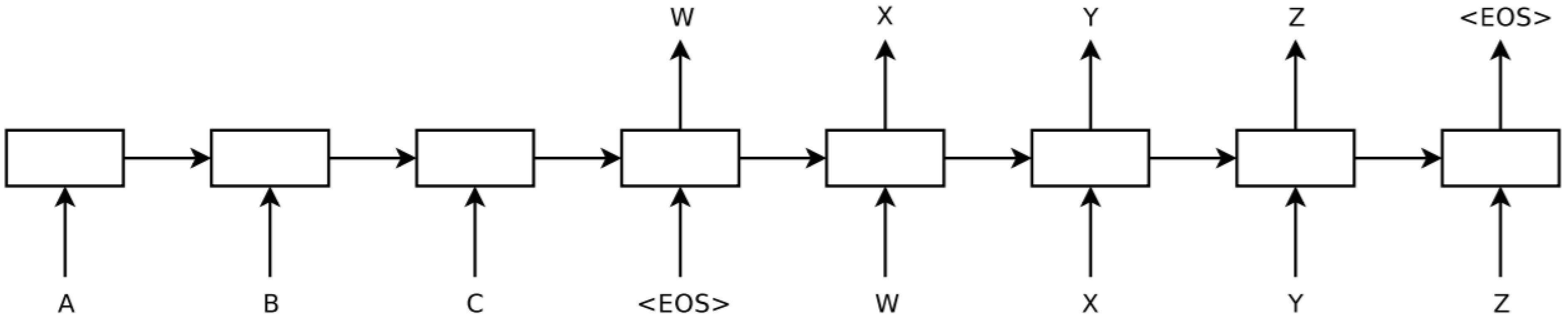
- Two RNNs for encoding and decoding of sequences, jointly trained
- Equations in the paper using gated recurrent unit
- They only looked at rescoring translation phrases, not generating



Literature Survey

Sequence to Sequence Learning with Neural Networks [s2s]
Sutskever et al. 2014
Keras, Numpy, Tensorflow

- Encoder-decoder with LSTM (pretty big architecture)
- Words are reversed in source sequence for better performance
- Left to right beam search decoder



Literature Survey

A Neural Conversational Model [chat]
Vinyals et al. 2015
Torch, Keras, Theano, Tensorflow

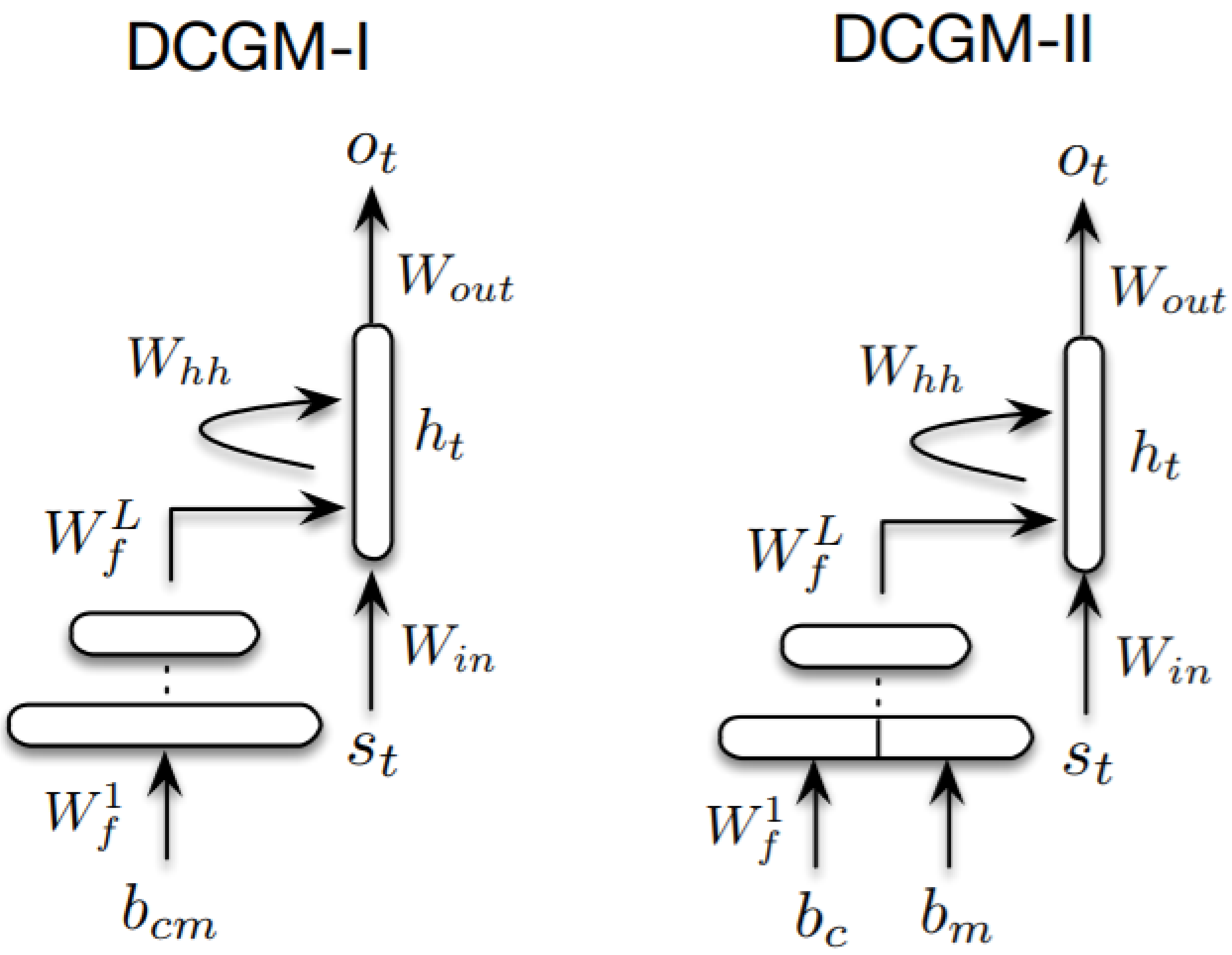
- IT helpdesk dataset and movie subtitles; Big architectures and big vocabs
- Input sequence is what has been conversed so far (context), output sequence is the reply
- Objective function optimized is not the actual objective achieved through human communication
- Problem mentioned is with the inconsistent answers (there is no personality) and with not being able to evaluate correctly

Literature Survey

[A Neural Network Approach to Context-Sensitive Generation of Conversational Responses](#) [chat]

Sordoni et al. 2015

- Encode past information, which is then decoded to promote responses
- Separate context from last message
- They use IR to generate more responses to a (c,m,r) triple based on bag of words
- They use a ton of features together with the neural network models to generate likely responses



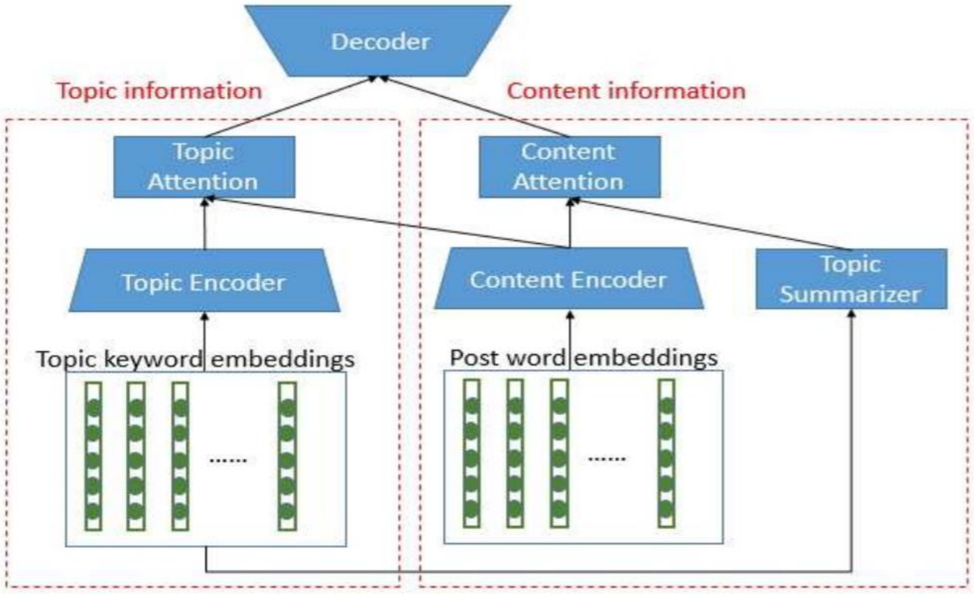
Literature Survey

Topic Aware Neural Response Generation [chat]

Xing et al. 2016

Theano

- Represent people's prior knowledge about the topic, and embed this into reply of seq2seq model with attention
- Two encoders with separate attention modules, one is bidirectional RNN, other is for topic words, then their attention is jointly fed into decoder
- Two encoders can affect each others attention, topic attn finds relevant info, content attn determines the content focus
- Topic word list obtained from twitter LDA model, they play the role of classification and association in response generation (better first words chosen)



Dataset

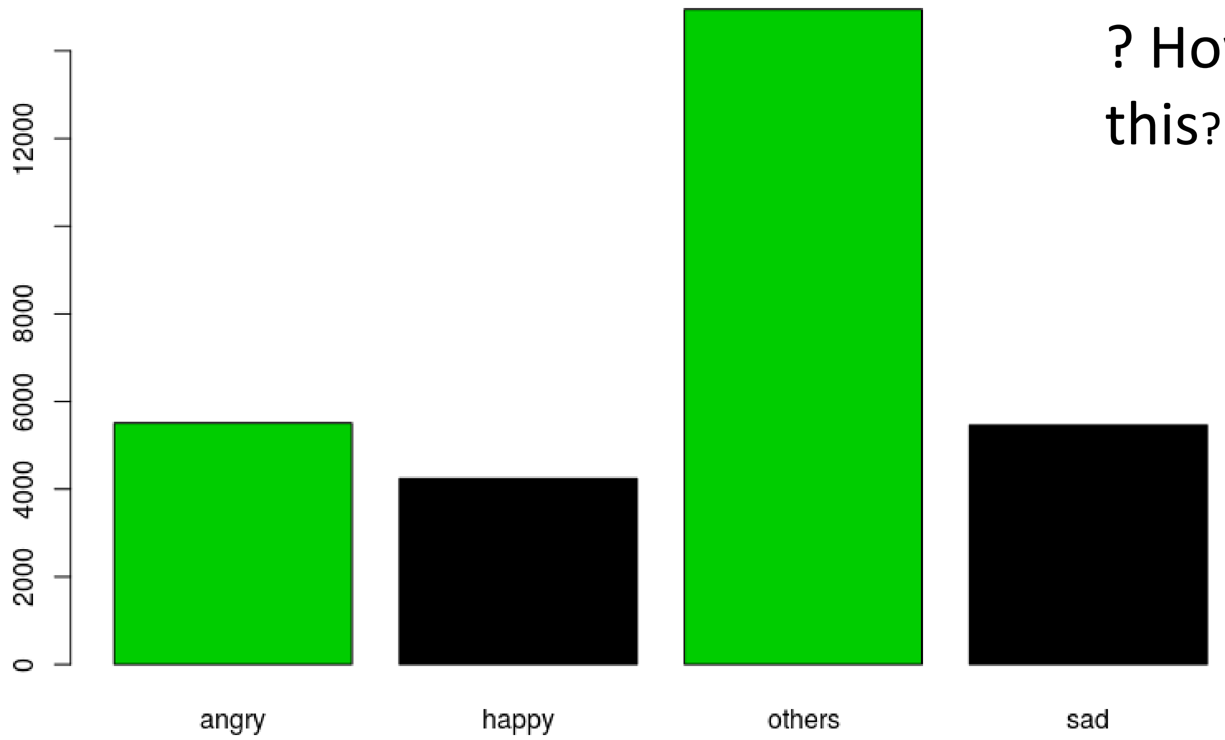
- There are 30160 message conversations labelled with their corresponding emotion scores.
- Obtained from $Emocontext\text{:}\ A\ Shared\ Task\ At\ SEMEVAL\ 2019$[1]
- Examples:
  - Turn 1: I am not feeling good.
  - Turn 2: Why, what happened?
  - Turn 1: I have been sick for one week.
  - Label: Sad ← **Only Categorical Label, No scores.**

  - Turn 1: For a computer pretending to be a human, you type too fast.
  - Turn 2: And are you pretending to be a lizard? Lol.
  - Turn 3: That's funny. Haha!
  - Label: Happy ← **Only Categorical Label, No scores.**

Dataset

? Only categorical
? How to improve this?

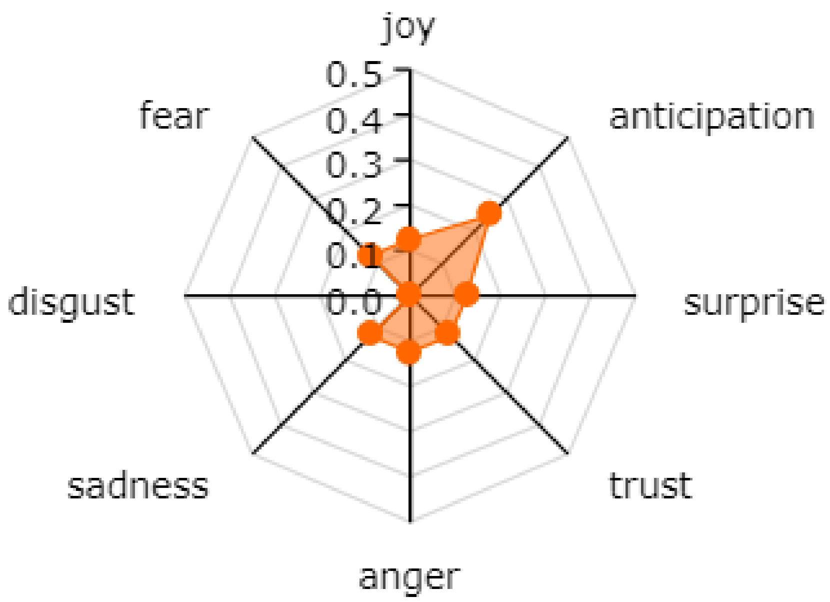
15k : Happy, Angry, Sad
15k: Others



??
"I am upset on how things are going and is mad about what John said the other day to Emilie.", there is 'sadness' and 'anger' both present to some extent. So it is wrong to label it with only one emotion.

Is only one emotion label per message is enough?

"Firenze also asks Harry if Harry would not know anyone who has been waiting for years to regain power and Harry realises that the mysterious figure was a weakened but still alive Tom ."



Source: Harshita Jhavar and Paramita Mirza. 2018. EMOFIEL: Mapping Emotions of Relationships in a Story. In WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.

Task 1: Developing a mapping between categorical and dimensional model of emotions

- Used the NRC Word−Emotion Association Lexicon (EmoLex)[2,3]
- Contains a list of words and their associations with eight basic emotions.
- Can analyze text in terms of categorical emotions instead of sentiment polarity (e.g., AFINN, SentiWordNet)
- Given emotion labels E ={anger, fear, anticipation, trust, surprise, sadness, joy and disgust} and a message conversation m containing n number of sentences, we compute the percentage of emotion e in the message, e ∈ E, as:

$$\text{percentage (m)} = \frac{\sum_{n}^{i=1} f_{ei}}{\sum_{n}^{i=1} \sum_{j \in E} f_{ji}}$$

where f_{ei} is the frequency of emotional words of label e in the ith message m

[2] Saif M. Mohammad and Peter D. Turney. 2010. *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. 26–34.*
[3] Saif M. Mohammad and Peter D. Turney. 2013. *Crowdsourcing a Word-Emotion Association Lexicon. 29, 3 (2013), 436–465*

Task 1: Developing a mapping between categorical and dimensional model of emotions

STEP 1: percentage (m) $= \frac{\sum_{n}^{i=1} f_{ei}}{\sum_{n}^{i=1} \sum_{j \in E} f_{ji}}$

STEP 2:

happy_score = joy + trust
anger_score = anger + disgust
sad_score = sadness
others_score = anticipation + surprise

BRINGING IT DOWN TO 4 LABELS AS GOLD CORPUS HAS 4 LABELS: Happy, Angry, Sad, Other(Neutral)

Task 1: Developing a mapping between categorical and dimensional model of emotions

EX:

Turn 1: *Are u blind? I am on my knee and I have the costliest ring in the world in front of u.*

Turn 2: *No, you're clearly in DENIAL !*

Turn 1: *No kill me pls*

STEP 1

NRC sentiment score: -0.58333333

Joy: 0.0
Anticipation: 0.0
Surprise: 0.0
Trust: 0.0
Anger: 0.0
Sadness: 0.5
Disgust: 0.0
Fear: 0.5

STEP 2

Happy Score: 0.0
Anger Score: 0.0
Sad Score: 0.5
Other Score: 0.0

Task 1: Developing a mapping between categorical and dimensional model of emotions

STEP 3:
- Russell and Mehrabian's Valence−Arousal−Dominance (VAD)[4] model is used
- Three fundamental emotional dimensions:
  - ✓ valence (the degree of pleasure or displeasure of an emotion),
  - ✓ arousal (level of mental activity, ranging from low engagement to ecstasy)
  - ✓ dominance (extent of control felt in a given situation).

- Used Jena Emotion Analysis System (JEmAS)[5] for analyzing the VAD score of each message
- Input: All sentences in the message

[5] Sven Buechel and Udo Hahn. 2016. Emotion Analysis as a Regression Problem - Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation.. In ECAI.
[6] Albert Mehrabian and James A Russell. 1974. An approach to environmental psychology. the MIT Press.

Task 1 Evaluation: Developing a mapping between categorical and dimensional model of emotions

- True categorical label = labelOf (max(happy.score, sad.score, angry.score, other.score)
  assigned by the emotion classifier for 72.6% cases only.

- This proves that the associated percentage of emotion share within different categorical emotions adds to the emotion information at the lexicon level.

- This information was lost when we only considered the emotion labels and not the emotion share of the scores.

Task 1: Developing a mapping between categorical and dimensional model of emotions

EX:

Turn 1: *Are u blind? I am on my knee and I have the costliest ring in the world in front of u.*

Turn 2: *No, you're clearly in DENIAL !*

Turn 1: *No kill me pls*

STEP 1

NRC sentiment score: -0.58333333

Joy: 0.0
Anticipation: 0.0
Surprise: 0.0
Trust: 0.0
Anger: 0.0
Sadness: 0.5
Disgust: 0.0
Fear: 0.5

STEP 2

Happy Score: 0.0
Anger Score: 0.0
Sad Score: 0.5
Other Score: 0.0

STEP 3

Valence: 0.11850
Arousal: -0.74600
Dominance: -0.02300

True categorical label matched with the labelOf(max(happy.score, sad.score, angry.score, other.score)) assigned by the emotion classifier for 72.6% cases.

Task 2: Emotion analysis of the user utterance

EX:

Turn 1: *Are u blind? I am on my knee and I have the costliest ring in the world in front of u.*

Turn 2: *No, you're clearly in DENIAL !*

Turn 1: *No kill me pls*

STEP 1

NRC sentiment
score: -0.58333333

Joy: 0.0
Anticipation: 0.0
Surprise: 0.0
Trust: 0.0
Anger: 0.0
Sadness: 0.5
Disgust: 0.0
Fear: 0.5

STEP 2

Happy Score: 0.0
Anger Score: 0.0
Sad Score: 0.5
Other Score: 0.0

STEP 3

Valence: 0.11850
Arousal: -0.74600
Dominance: -0.02300

STEP 4

Subject_Characters
Actions_With_Subject_Characters

Object_Characters
Actions_With_Object_Characters

Dependency Parsing

Task 2 Evaluation : Emotion analysis of the user utterance

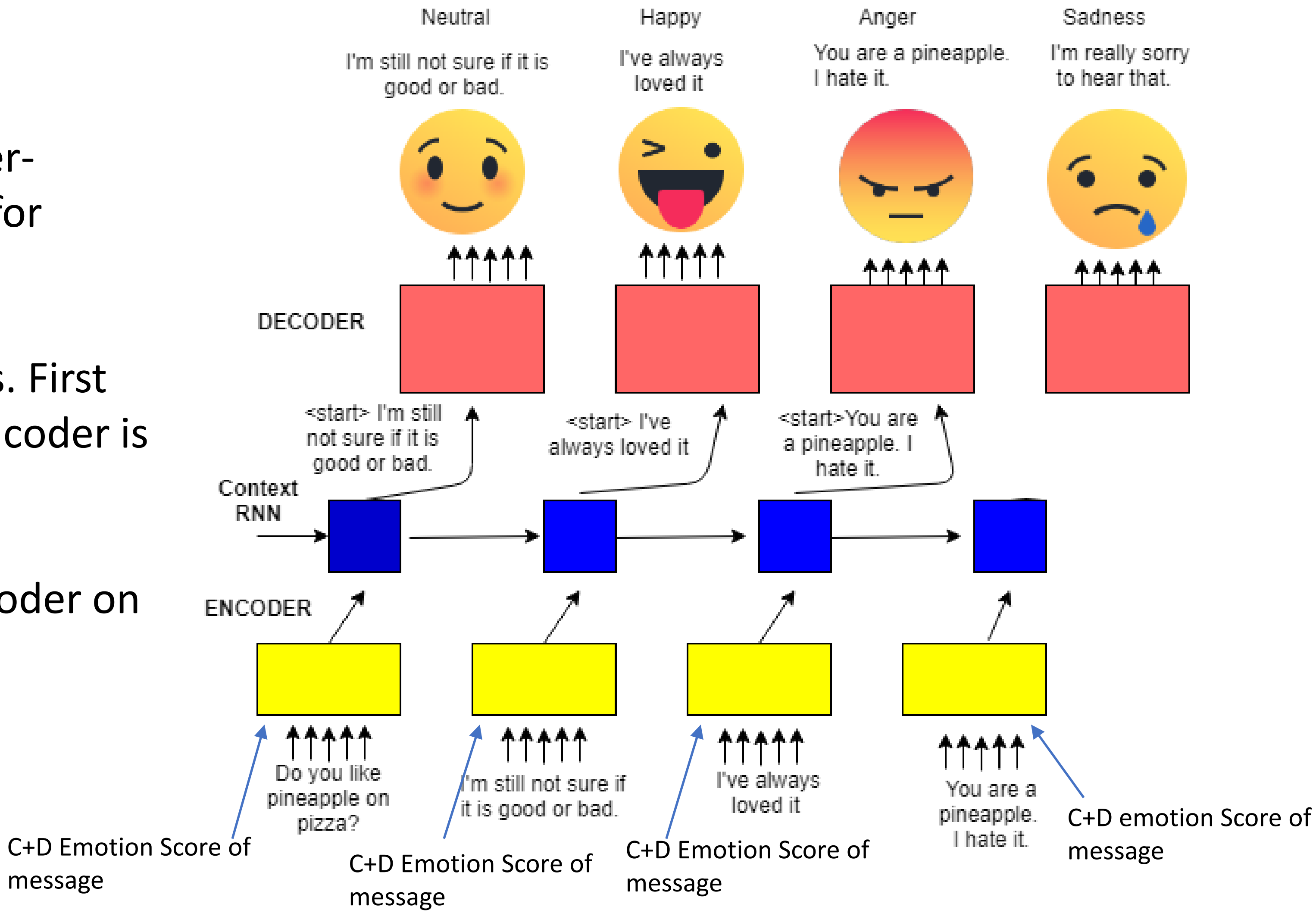
- Trained a classifier on the dataset prepared in Goal 1
- Model: Basic LSTM
  - Categorical Cross Entropy as Loss Function
  - Rmsprop as optimizer
  - Sigmoid as activation

Feature	C.V. Accuracy	Coda Lab Score
Label $\sim Turn1 + Turn2 + Turn3$	0.8342	0.6064
Label $\sim Turn1 + Turn2 + Turn3 + Sentiment\_Score$	0.7921	0.5912
Label $\sim Turn1 + Turn2 + Turn3 + Happy\_Score+$ Sad_Score + Anger_Score + Others_Score + Sentiment_Score	0.8672	0.6978
Label $\sim Turn1 + Turn2 + Turn3 + Valence\_Score+$ Arousal_Score + Dominance_Score + Sentiment_Score	0.8578	0.6624
Label $\sim Happy\_Score + Sad\_Score + Anger\_Score + Others\_Score$	0.7341	0.5525
Label $\sim Valence\_Score + Arousal\_Score + Dominance\_Score$	0.7520	0.5642
Label $\sim Turn1 + Turn2 + Turn3 + Sentiment\_Score$ + Subject_Actions + Object_Actions	0.5238	0.4016
Label $\sim Turn1 + Turn2 + Turn3 + Happy\_Score+$ Sad_Score + Anger_Score + Others_Score + Subject_Character + Subject_Actions + Object_Character + Object_Actions + Sentiment_Score	0.8186	0.5914
Label $\sim Turn1 + Turn2 + Turn3 + Valence\_Score+$ Arousal_Score + Dominance_Score + Subject_Character + Subject_Actions + Object_Character + Object_Actions + Sentiment_Score	0.8512	0.6418
Label $\sim Turn1 + Turn2 + Turn3 + Happy\_Score+$ Sad_Score + Anger_Score + Others_Score + Valence_Score + Arousal_Score + Dominance_Score	0.9516	0.7427
Label $\sim Turn1 + Turn2 + Turn3 + Happy\_Score+$ Sad_Score + Anger_Score + Others_Score + Valence_Score + Arousal_Score + Dominance_Score + Subject_Character + Subject_Actions + Object_Character + Object_Actions + Sentiment_Score	0.8414	0.6819

Task 3: Generate emotional responses from chatbot

Model Details:

- Hierarchical Recurrent Encoder-Decoder (HRED) architecture for handling deep dialog context
- Multilayer RNN with GRU cells. First layer of the utterance-level encoder is always bidirectional.
- Thought vector is fed into decoder on each decoding step.



Task 3: Generate emotional responses from chatbot

Decoding Details

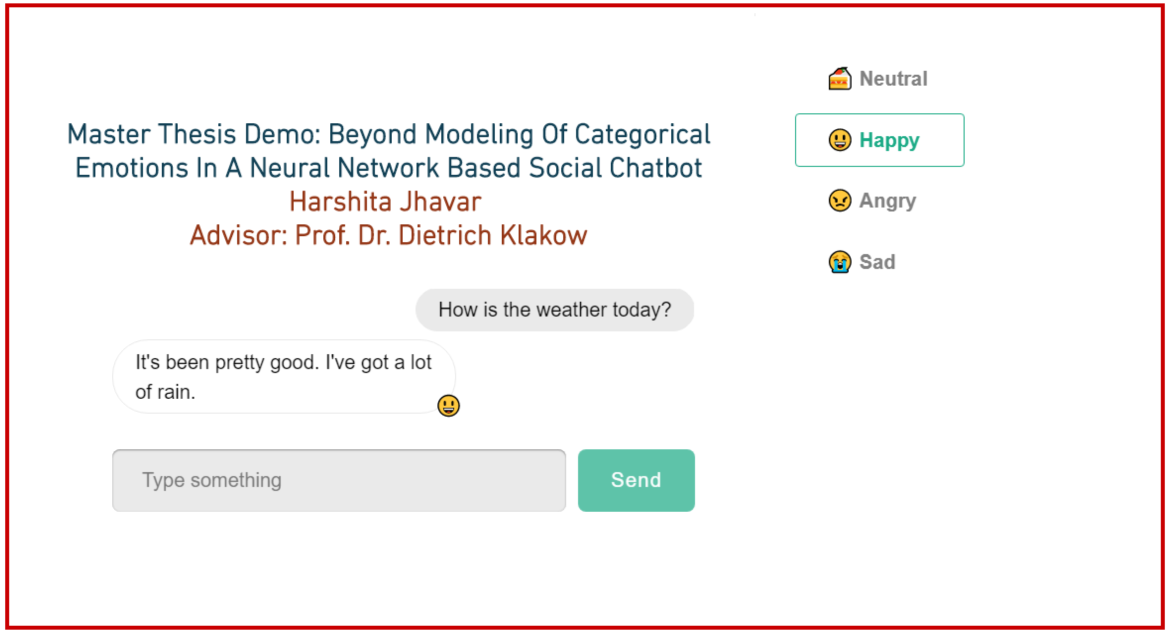
- Decoder is conditioned on emotion label.
- The model is trained with **context size 3** where the encoded sequence contains **30 tokens or less** and the decoded sequence contains **32 tokens or less**.
- Initialized using word+emotion_scores to vector model trained on the corpus developed in goal 1 and goal 2.
- Both encoder and decoder contain **2 GRU layers** with **512 hidden units** each.

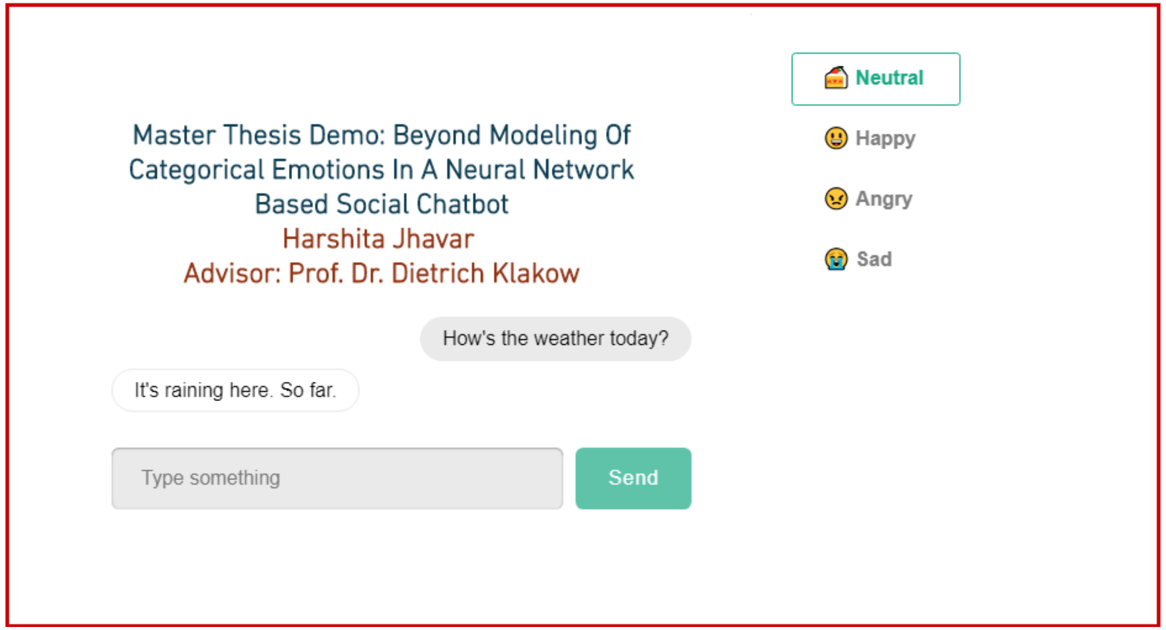
Model Details:

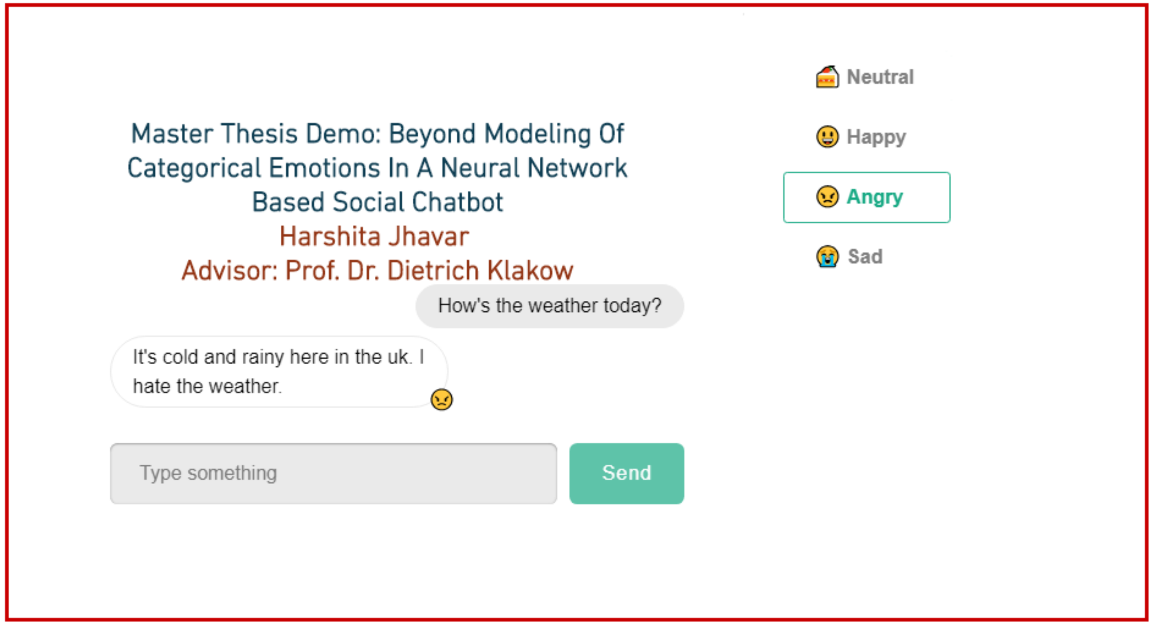
- Response generation through sampling-reranking algorithm.
- Reranking of the generated candidates is performed according to the log-likelihood.

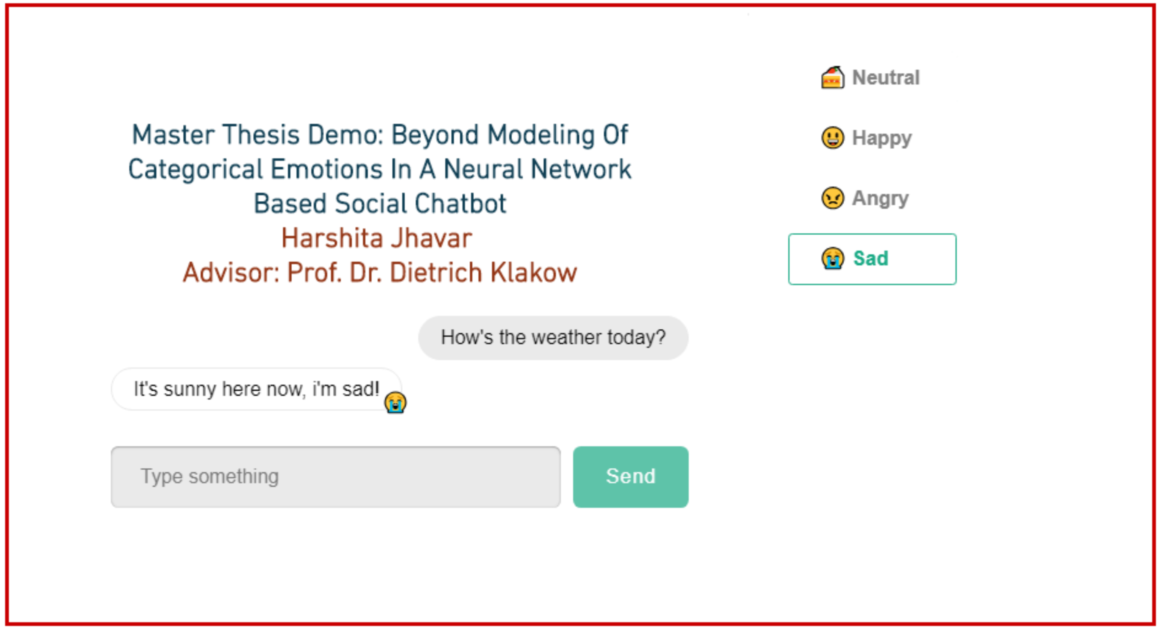
Model architecture inspired from the replika team model architecture. However, my thesis made it possible to add emotion scores and semantic scores to be incorporated in the model. Also, the dataset is completely different than the replica team data source.
https://github.com/lukalabs/cakechat

Task 3: Generate emotional responses from chatbot









Task 3 Evaluation : Generate emotional responses from chatbot

- Human evaluation
- 30 participants
- Google form

Beyond Modeling of Categorical Emotions in a Neural Network Based Social Chatbot

Please choose the appropriate emotion category (Only One) corresponding to the responses to the questions below and also, choose in last two columns if the response makes sense to the asked question.

Choose emotion from: happy, anger, sad, other *

	Happy	Angry	Sad	Neutral/ Other	Response doesn't make sense	Response makes sense.
Q: What do you think about me? R: I don't know.. you are perfect.	☐	☐	☐	☐	☐	☐
Q: Why are you so stressed out? R: I am just not feeling well.	☐	☐	☐	☐	☐	☐
Q: How are you doing? R: Why are you so mad at me?	☐	☐	☐	☐	☐	☐
Q: Do you like pineapple on pizza? R: I don't like pineapple.	☐	☐	☐	☐	☐	☐
Q: Do you like your President? R: Yes, I like her.	☐	☐	☐	☐	☐	☐
Q: Did the rain stop? R: It was a joke.	☐	☐	☐	☐	☐	☐
Q: How was the food in the event? R: I have no idea.. it was horrible.	☐	☐	☐	☐	☐	☐
Q: Was the movie good? R: Yeah, it was scary.	☐	☐	☐	☐	☐	☐
Q: Did you win the game? R: No, I lost the first game.	☐	☐	☐	☐	☐	☐

Task 3: Generate emotional responses from chatbot

Record sheet from one participant

EVALUATING THE EMOTION INVOLVED IN THE GENERATION OF RESPONSE

30 Participants
30 Queries + Replies
Google Form Link

	Happy	Anger	Sad	Other
Happy	6	0	0	1
Anger	0	5	2	0
Sad	0	1	6	0
Other	1	1	1	6

	True Positive	False Positive	True Negative	False Negative	Precision	Recall	Accuracy	F1 Score
Happy	6	1	22	1	0.85	0.85	0.93	0.85
Anger	5	2	21	2	0.71	0.71	0.87	0.71
Sad	6	1	20	3	0.85	0.67	0.87	0.75
Other	6	2	21	1	0.67	0.85	0.87	0.75

Task 3 Evaluation : Generate emotional responses from chatbot

EVALUATING THE EMOTION INVOLVED IN THE GENERATION OF RESPONSE

Conclusion for 30 participants

	Average Precision	Average Recall	Average Accuracy	Average F1-Score
Happy	0.874	0.889	0.944	0.8814
Anger	0.801	0.821	0.882	0.8108
Sad	0.88	0.861	0.929	0.8703
Other	0.752	0.791	0.853	0.7710

Task 3 Evaluation : Generate emotional responses from chatbot

EVALUATING THE QUALITY OF RESPONSE

Data from one participant

30 Participants
30 Queries + Replies
Google Form Link

	Response Makes Sense	Response Doesn't Make Sense
Happy	7	0
Anger	6	1
Sad	7	0
Other	7	2

Conclusion for 30 participants

	Response Makes Sense	Response Doesn't Make Sense
Happy	98.3%	1.7%
Anger	96%	4%
Sad	98%	2%
Other	95.2%	4.8%

Contribution

1. Developed a mapping between the dimensional and categorical model of emotions.

2. Built a corpus labeled with the corresponding emotion scores for 30,160 message instances.

3. More than one categorical emotion present in a message also got scored according to the share of the emotion content present at the level of the lexicons used in the messages.

4. Performed emotion analysis of the user utterance.

5. Took as an input, the better emotionally informed corpus developed in goal 1 to the chatbot to generate emotional responses accordingly. With human evaluation, concluded that more than 95% users found that the responses generated were sensible to the context of the question asked.

6. Human evaluation also informed that the emotional responses generated corresponding to the emotion category chosen was reflected as an emotion in the generated responses with an average accuracy for each categorical emotion for around
88%.

Future Work

- Improve on context personalization, domain understanding and abstraction on the existing emotional bot.

- Develop better strategies to perform evaluation for chatbots.

- Extend the emotional conversational multimodal to speech output with emotional tones. A chatbot with emotional tones of surprise, sadness will be interesting to have.

- Include Topic Based Response Generation i.e. incorporating emotions for domain specific chatbot response generation.

- Incorporating mood of the user to decide independently the mood of the chatbot.

- Incorporating persona in a chatbot. Ex: Your chatbot could be Yoda from Star Wars. May the force be with you!

THANKYOU FOR YOUR KIND ATTENTION!