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

by

Harshita Jhavar

Matriculation Number: **2566267**

Master Thesis

Lab for Speech And Signal Processing
Faculty of Computational Linguistics and Phonetics
And Department of Computer Science
Saarland University, Germany

Supervisor
Prof. Dr. Dietrich Klakow

Advisor
Prof. Dr. Dietrich Klakow

Reviewers
Prof. Dr. Dietrich Klakow
Prof. Dr. Vera Demberg

February 22, 2019

UNIVERSITÄT
DES
SAARLANDES

Declaration of Authorship

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Datum/Date: _____

Unterschrift/Signature: _____

Erklärung

Ich erkläre hiermit, dass die vorliegende Arbeit mit der elektronischen Version übereinstimmt.

Statement

I hereby confirm the congruence of the contents of the printed data and the electronic version of the thesis.

Saarbrücken, Datum/Date: _____

Unterschrift/Signature: _____

SAARLAND UNIVERSITY, GERMANY
Lab for Speech And Signal Processing
And Department of Computer Science

Abstract

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

by Harshita Jhavar

Conversation modeling is an important task in natural language understanding and machine intelligence. One of the key improvements for conversation modeling is to make the experience with the AI agent more human like. Generating human-like responses improves the overall user experience. Motivated with this very thought, the work of my Master thesis revolves around three important tasks. In Task 1, I developed a cross-model emotion mapping on message conversation corpus. I assigned the corresponding categorical emotion scores for each category(Happy, Sad, Anger, Other) and dimensional emotion scores(Valence, Arousal, Dominance) corresponding to each message. The categorical scores define a percentage of each category of emotion shared in the message. In Task 2, I focused on emotion analysis of the user utterance by training an emotion classifier on the corpus developed in Task 1. In Task 3, I used the developed corpus in Task 1 and conclusions from the feature engineering performed in Task 2 to finally feed the chatbot's Hierarchical Recurrent Encoder-Decoder based architecture with the right set of features. I performed evaluation for the classifier developed in Task 2 on CodaLab Testset for EmoContext: Semeval task 2019[1] to achieve the best accuracy of around 74%. I performed human evaluation for Task 3 with 30 participants to conclude that more than 95% users found that the responses generated was sensible to the context of the question asked. Another conclusion from the human evaluation was 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%.

[1]https://www.humanizing-ai.com/emocontext.html

Acknowledgements

I would like to thank Prof. Dr. Dietrich Klakow for his guidance and support during the entire course of my thesis completion and being contant reviewer of my work. Dietrich's emphasis on practices like documenting ideas on a daily basis, striking a balance between where to push to get results and where to backtrack to explore another path and its influence on the thesis productivity further honed my confidence to complete my thesis. There has always been a perfect balance between guidance and freedom under his tutelage.

Many thanks to Xiaoyu Shen from our group who has helped and encouraged me a lot with his suggestions on improving for more robust results. Thanks to Prof. Dr. Vera Demberg for agreeing to be reviewer for my thesis.

At last, I want to thank my parents, my sisters and my friends who have always given their unconditional support and love to me. It was their constant emphasis on being always strategic, focussed and well planned to complete my master thesis in parallel to new onboarding learnings at Microsoft Prague.

Content

List of Figures

List of Tables

Dedicated to my grandmother, Late Smt. Kamla Maheshwari

Chapter 1

Introduction

1.1 The Family of Chatbots

Early conversation systems like 'Eliza' (Weizenbaum, 1966), Parry (Colby, 197), and Alice (Wallace, 2009) were designed on hand-crafted rules and had controlled scope. A lot of domain specific research has been done for building chatbots for domain specific goals like 'Building a system for reserving airlines tickets' or 'Booking a table in a restaurant'. These task completion chatbots are generally driven by data driven machine learning based approaches.

A tremendous amount of progress has been made to develop intelligent personal assistants like Microsoft's Cortana, Apple's Siri, Google's Google Assistant, Amazon's Alexa etc. These bots have a wider scope of answering to user queries. These bots are also capable of giving reminder and recommendation assistance without receiving requests from the user. These applications are extremely important as people have started depending to manage their schedules using these applications.

On the other hand, there exist social chatbots like Microsoft's social chatbot XiaoIce developed by Shum et. al 2018 [2], where the primary goal of this social bot is not to solve queries from the users but to be more like a companion with the users. Due to the advancements in technologies for natural language understanding, speech recognition, information retrieval, cognitive science, computer vision and related fields, social chatbots have gained more popularity. Social bots which generate emotional responses will be the primary focus of my master thesis. Developing an emotion labelled corpus which defines a mapping between the categorical and dimensional emotions is another major contribution from my thesis work.

1.2 Thesis Roadmap

After discussing the motivation for my thesis in chapter 2, I briefly discuss the literature survey for emotional social bots in chapter 3 which include the state of art for generating emotional responses in a social chatbot and a study of the existing set of relevant datasets. In chapter 5, 6, 7; I discuss the three important tasks performed during my Master thesis work and the results of their evaluation. Finally, I discuss the conclusion of my thesis and future work in Chapter 8 with mentions of relevant references in Bibliography in the end.

Chapter 2

Motivation

2.1 Why emotions in chatbot are important?

AI these days is getting more human-centric than being machine centric. Computing for better human experiences is state of art driving the different tools and technologies. Thus, it is critical to align technology with human's multi sensory capabilities and correspondingly deal with human's multimodal information. While chatbots have been task specific like insurance chatbot, financial chatbot, hotel chatbot, with/ without domain knowledge, with/ without contextualized personalization, abstraction; there has also been motivation towards incorporating emotions in the replies generated from chatbot. While it is important to focus on the quality of the task performed by the chatbot but it is also important to focus on the experience of the user. Our daily conversations involve emotions. We choose words in our conversations which we think are appropriate to express our emotions. An emotionally rich AI model makes the user experience much more appealing and enjoyable than template based replies. Thus, improving in terms of building an emotional bot is very important.

2.2 What are the challenges to built an emotionally rich chatbot?

The challenges are as listed below:

- The corpus used for training for such chatbots usually have only true labels marked to it. For example: Labels like 'Anger, Sad' etc. However, in real life, we use the same set of words to express different quality of emotions. Every sentence shares a percentage of emotions shared and thus, it is not fully correct to consider a

message with just one emotion score. It is important to consider the percentage of emotion share in a particular message and pass that information to the chatbot in order to improve the quality of the responses.

- Dimensional model of emotions which involves scoring for Valence, Arousal and Dominance in a particular message. Valence is the degree of pleasure or displeasure of an emotion, arousal is the level of mental activity, ranging from low engagement to ecstasy and dominance is extent of control felt in a given situation.

- There exists no such cross mapping which maps categorical emotions to dimensional emotions for message conversations to my knowledge. So this kind of corpus has to be built first which describes Task 1 and Task 2 of my thesis as discussed in the next chapters.

Chapter 3

Literature Survey on Social Chatbots and Relevant Datasets

3.1 State of art for incorporating emotions in social chatbots

Talking about the state of the art for building emotional chatbots in framework of neural network, here is a discussion about some recent papers from the field. In Zhou et al. 2018 [3], an emotional chatting machine is proposed that can generate emotional responses not only in content but also in emotional consistency. According to the authors, there has been no prior work for large scale emotional conversation generation with architecture inspired from neural network in past. A seq2seq framework is used with emotion category embedding, internal implicit emotion memory, and external explicit memory. P(Y|X,e), where e is one of the 6 emotion categories, Y is the response and X is the query input which is embedded and is fed into the decoder. Internal memory: captures emotion dynamics where each emotion is decaying during decoding because it is read and written (by the GRU) at each step to the memory. External memory allows the model to choose between words from a generic or an emotion vocab (separate softmaxes). For the purpose of regularization, emotion state in internal memory should decay to zero at the end of decoding; there is another term for constraining the external memory. Emotion category annotation is obtained with Bi-LSTM emotion classifier (accuracy 62.3 %). It is also concluded that the ECM model obtains better perplexity (without external memory) and emotional accuracy and better human rating than base seq2seq architecture. Thus, this paper discusses three new mechanisms that respectively (1) Models the high-level abstraction of emotion expressions by embedding emotion categories, (2) Captures the

change of implicit internal emotion states, and (3) Uses explicit emotion expressions with an external emotion vocabulary. In Gupta et al. 2018 [4], a different deep learning based approach called Sentiment and Semantic LSTM (SS-LSTM) is discussed to detect emotions in textual conversations. The approach combines sentiment and semantic features from user utterance using sentiment and semantic word embeddings and gloVe embeddings respectively and do not require any hand-crafted features. The evaluation on real world textual conversation has proven that the approach outperforms CNN and LSTM baselines, in addition to other Machine Learning baselines. An industrial product based on thsi architecture is discussed in Shum et al. 2018 [2] in which Microsoft's social chatbot XiaoIce has been used to perform case studies for various technologies to build social chatbots. XiaoIce can dynamically recognise the categorical set of emotions from user responses and engages the user for long conversations. In Wang et al. 2018 [5], instead of the above two different neural network based architectures, Generative Adversarial Neural Networks are implemented to generate different sentiment labelled sentences without supervision. There are multiple set of generators with one multi-class discriminator for this task. For the purpose of evaluation, they have defined 4 measures to define the quality of the sentences: Fluency, Novelty, Diversity, Intelligibility. For fluency, SRILM was used which returns the perplexity of the generated sentence. For novelty, it tests if the generator simply copies the sentence from the corpus or not. For diversity, maximum Jaccard similarity calculated between each generated sentence and the rest of the sentences is used as a measure. For intelligibility, human evaluation was performed on the generated text by three humans by asking them to rate 100 generated sentences and rate it between 1 to 5. However, this paper only deals with sentiments(positive or negative) but not with emotions which was the case with the first two papers mentioned above. The reason to mention the last paper was to summarize the various possible evaluation strategies.

3.2 State of the art for relevant datasets

In Zhou et al. 2018 [6], a huge corpus is built by preprocessing Twitter conversations by taking the emoticons as labels of emotions on the different tweets. Autoencoders training are performed on these conversations with several conditional variations to control the generated text emotion. They claim in their quantitative and qualitative analyses that the proposed models can successfully generate high quality conversation responses in accordance with designated emotions. They evaluated the system with human evaluation and automatic emotion classifier they built to verify the correctness of the emotion generated in the response. However, these responses are only with one category and doesn't include the scope of generating responses in more than one category

FIGURE 3.1: Examples of Labeled Message Instances From The Dataset

- 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.

in a single response. Also, the quality of the responses is not upto the mark. In Li et al. 2017 [7], a dialog dataset has been introduced where the dialogs are in English. Short dialogs on specific topics form the corpus (average dialog is 8 turns). Each utterance is labeled as one of four dialog acts: inform, questions, directives, commisive. Each utterance is labeled as one of 7 emotion categories: anger, disgust, fear, happiness, sadness, surprise, other. Dialogs usually follow some pattern along the four dialog acts, like question-inform bi-turn dialog flow. 83% of dialogs falls into the "other" emotion category.

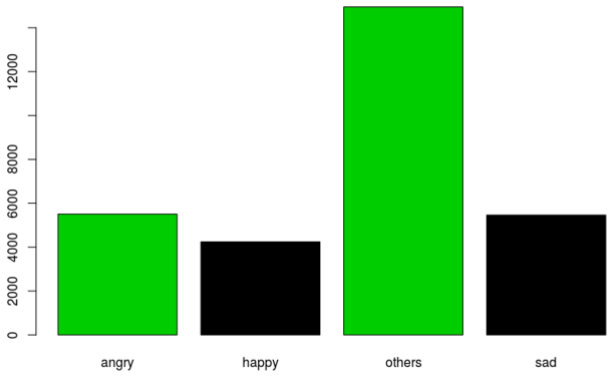
3.3 Thesis Dataset

In follow-up to the work discussed in Gupta et al. 2018 [4] by Microsoft AI and Research team, they have organised a shared task competition called as SemEval19 Task: EmoContext[1] . They have released a dataset which comprises of 15K records for twitter conversation for three emotion categories i.e., Happy, Sad and Angry. It also contains 15K records not belonging to any of the aforementioned emotion classes and is mentioned with the class label 'Other' as shown in Figure 3.2. Examples of message instance as recorded in the dataset is given in Figure 3.1.

However, this dataset has categorical variables only in form of labels. There is no categorical variable scores assigned to the messages nor there is any dimensional score assigned to the data. Also, the dataset assumes that there is only one type of emotion present in the data while in reality, there is always a mix of emotions present. For example: In a sentence like, "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. We will discuss this point further in description for Task 1 in Chapter 5.

[1] https://competitions.codalab.org/competitions/19790learn$_t$he$_d$etails

FIGURE 3.2: Distribution of Categorical Variables In The Dataset



Chapter 4

Problem Statement: Three Goals

The problem statement for my Master Thesis can be divided into three tasks for the following three goals: **Goal 1: Develop a mapping between the dimensional and categorical model of emotions** and building a corpus labeled with the corresponding emotion scores for the two models of emotion theory: categorical and dimensional. Another important part of this task is to perform: **Assignment of categorical emotion scores in form of percentage share of different categorical emotions in a particular message.** For example: If words used in a sentence could imply sadness and anger together, then defining the percentage of sadness and anger expressed in that particular message and declaring the category with the largest percentage as the emotion label for that message.

Figure 4.1: Task Flow Checkpoints for My Master Thesis

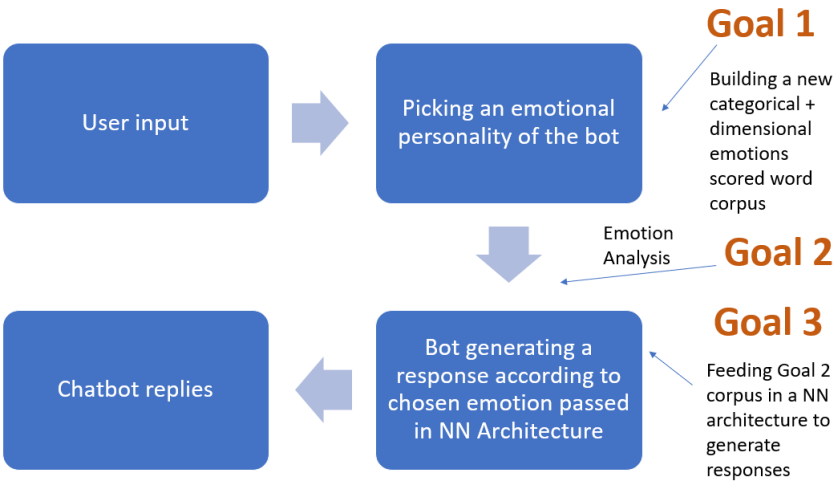
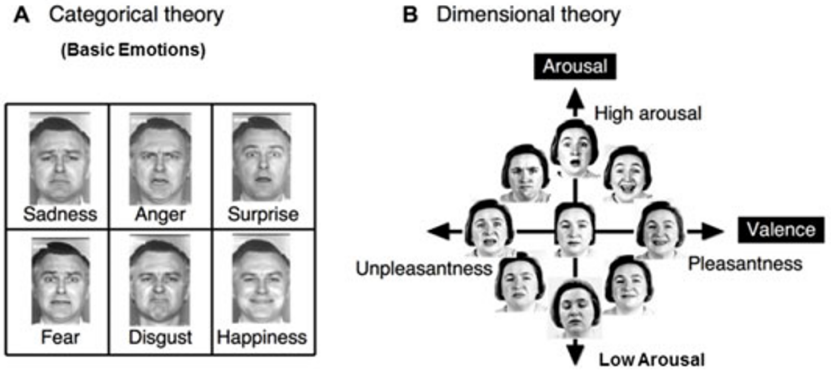


FIGURE 4.2: 2 Models of Emotion Theory: Source: Matsuda et al(2013)[1]



Goal 2: Emotion analysis of the user utterance. This goal focuses on training a classifier for the emotion analysis of the utterances by the user. The improved corpus developed in Goal 1 which is more emotionally informed about each message instance recorded in it is used for this task.

Goal 3: Generate emotional responses from the chatbot as per the chatbot emotion personality input by the user. Ex: Happy, Sad, Angry, Other/Neutral.

4.1 Elaborating Goal 1

All the existing work for generating emotional responses in chatbots only incorporates explicitly mentioned categorical set of emotions. These categorical set of emotions are derived from the Ekman's set of emotion categories: Happy, Sad, Anger, Disgust, Fear. The Dimensional model of emotion theory is also known as VAD (Valence, Arousal, Dominance) model of emotion theory as shown in Figure 4.2. It takes into account the different intensity or the amount of emotion present in a response. This can also be described as incorporating dimensional model of emotions alongside the categorical set of emotions as per the basics of the emotion theory. The idea is to score each of 30,160 message conversation in the data corpus with dimensional scores and categorical scores. This will lead to the establishment of cross mapping between the two models of emotion theory.

4.2 Elaborating Goal 2

This goal deals with emotion analysis of the user utterance. Once goal 1 is accomplished, we need to train a classifier in order to perform emotion analysis for the user utterances

and assign the right emotion label to the message. The messages are labeled in turns of user1, user2, user1 utterances and are labeled with the corresponding emotion label which are labels for the context. Lexicon based emotion classifier are not enough as they fail in instances when the sentence contains phrases like, "not happy" as 'not' will lead to sadness or anger while 'happy' will lead to 'joy'. More than one lexicon can correspond to more than one emotion. For example: A word like 'Mad' can mean 'Anger' ("I am mad at you.") or 'Happiness' ("I was madly laughing.") depending on the context it is used. A user response like, "I did not win the lottery, damn it!" contains not only categorical emotion 'sadness' in it but there is some amount of categorical emotion 'anger' also included in it. Again, for example, in sentences like, 'I hate you' vs 'I don't like you', there is different share of 'anger' and 'sadness' in it. Thus, it becomes important to talk about the percentage of share of emotion in the word rather than just one categorical label. This is making the chatbot emotionally well learned. So, a lexicon based emotion classifier also fails in such instances. Thus, while it is necessary to incorporate the entire set of sentences in a message as input to take the context of the sentence into consideration, it becomes also necessary to include the right set of emotion feature vectors along. The goal 2 will be completed with the development of a robust emotion classifier.

4.3 Elaborating Goal 3

The third goal is to generate emotional responses from a social chatbot. The user shall be able to choose the emotional state of the bot among the set of Happy, Sad, Angry, Neutral(Other). So, I will train a LSTM model for chatbot and will input the corpus prepared in Goal 1 and improved feature set concluded in Goal 2 to generate appropriate responses. The details are discussed in the follow up chapters.

Chapter 5

Goal 1: Implementation and Evaluation

5.1 Definition:

Developing a cross-model emotion mapping on message conversation corpus and assign the corresponding categorical emotion scores for each category(Happy, Sad, Anger, Other) and dimensional emotion scores(Valence, Arousal, Dominance) corresponding to each message. The categorical scores define a percentage of each category of emotions shared in the message.

5.2 Categorical Model

One can use 'DeepMojiClassifier' [1], developed by MIT, however, this classifier generates emojies as a result and not the corresponding label. Also, the granularity of this classifier was at the level of sentence. I needed to train a Lexicon Based Classifier to develop scores for a message instance for the categorical model. I used *NRC Word-Emotion Association Lexicon (EmoLex)* [8, 9], which contains a list of words and their associations with eight basic emotions. This lexicon served my purpose of analyzing text in terms of categorical emotions instead of sentiment polarity (e.g., AFINN [10], SentiWordNet [11]) and with richer emotion categories compared with the dataset released for the WASSA-2017 Shared Task on Emotion Intensity [12, 13]. SentiWordNet 3.0 [11] has only sentiments as its categorical labels (positive, negative, neutral), however I am more interested in mapping emotions and not the sentiments. The Shared Task on Emotion Intensity [**?**] takes only four emotion categories (joy, anger, fear, sadness) while when I analyzed the *NRC Word-Emotion Association Lexicon (EmoLex)* [8, 9] to find that it contains a list

[1] https://github.com/bfelbo/DeepMoji

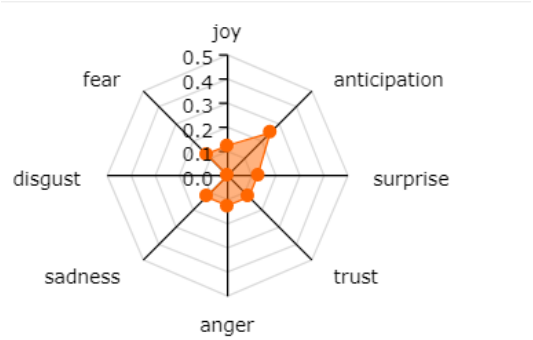


FIGURE 5.1: Categorical Emotion mapping example for a sentence

FIGURE 5.2: Dimensional emotion mapping example

of words and their associations with eight basic emotions and two sentiments (*negative* and *positive*), to analyze the emotion of a message. So, I chose NRC EmoLex dataset for categorical emotion mapping. In Jhavar et. al 2018[14], I have worked on performing a similar analysis on scenes in novels to analyze emotions in fictional characters. These, however, are message instances and the corpus is prepared for a chatbot.

Given emotion labels E ={*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy* and *disgust*} and a message m containing n number of sentences in a message, I compute the percentage of emotion e in the scene, $e \in E$, as:

$$\text{perc}_e(s) = \frac{\sum_n^{i=1} f_{ei}}{\sum_n^{i=1} \sum_{j \in E} f_{ji}} \tag{5.1}$$

where f_{ei} is the frequency of emotional words of label e in the i^{th} sentence of m.

An illustrative example of emotion mapping with both categorical and dimensional models for a sentence is given in Figure 5.1 and Figure 5.3.

Sentence	Joy	Trust	Anger	Sadness	Disgust	Fear	Valence	Arousal	Dominance
He will be embarrassed if people find out that they are related to them.	0	0.11	0.22	0.33	0.11	0.11	0.85	-1.09	0.55

TABLE 5.1: Example of scoring on a single sentence

5.3 Dimensional Model

Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model [15] is among the most commonly used dimensional approach, in which the emotional states can be described relative to 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) and *dominance* (extent of control felt in a given situation). I utilized Jena Emotion Analysis System (JEmAS) [16] for analyzing the VAD score of each scene, giving as input all sentences in the scene.

An example on how the hybrid set of scores are assigned for categorical and dimensional model of emotions is given in Table 5.1. Thus, the Goal 1 of developing a corpus of around 30,160 message instances, labelled with the corresponding emotion scores from the two models: categorical and dimensional, is completed successfully.

5.4 Narrowing to 4 categorical labels

An illustrative example of emotion scoring with both categorical and dimensional models for a message is given in Figure 5.1. However, I decided to assign four emotional personalities to the chatbot: Happy, Sad, Anger, Neutral(Other) following the class labels followed by the EmoContext: Semeval task 2019[2] . To obtain the scores for the following 4 classes, I used the following 4 rules and example of their implementation is shown in Figure 5.3.
happy.score = joy.score + trust.score
anger.score = anger.score + disgust.score
sad.score = sadness.score
others.score = anticipation.score + surprise.score
The scores of Valence, Arousal And Dominance remain unaffected.

[2]https://www.humanizing-ai.com/emocontext.html

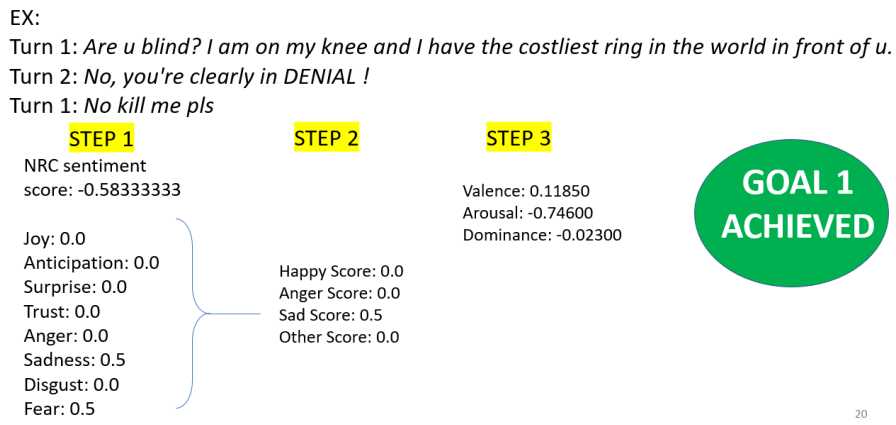


FIGURE 5.3: Emotion mapping scores on a message: Categorical and Dimensional

5.5 Evaluation

It was found that the true categorical label which was originally present in the dataset matched with the 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.

Chapter 6

Goal 2: Implementation and Evaluation

6.1 Definition

This goal focuses on training a classifier for the emotion analysis of the utterances by the user. The improved corpus developed in Goal 1 which is more emotionally informed with emotion scores for categorical and dimensional set of emotions for each message instance recorded in it, is used for this task.

6.2 Extending The Feature Set

From goal 1, the feature set comprised of the *User1.message*, *User2.message*, *User1.message.Reply*, *Happy.Score*, *Sad.Score*, *Anger.Score*, *Other.Score*, *Valence.Score*, *Dominance.Score*, *Arousal.Score*, *Sentiment.Score*.
However, since the actions associated to the subjects and objects in the sentence expresses emotions involved in a sentence, I chose to experiment with increasing the feature set using dependency parsing. I ran a dependency parser to parse Subject and its associated actions, Object and its associated actions.

6.3 Model Of The Emotion Classifier

I trained a basic LSTM classifier model where I used categorical cross entropy as the loss function. Rmsprop optimizer was used in the model which was built in R. Sigmoid was used as activation function. The input as w2v format alongside the cross validation accuracy is discussed in the follow up section.

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

TABLE 6.1: Feature Engineering Performed On The Classifier

6.4 Evaluation: Feature Engineering experiments

Table 6.1 shows the cross validation accuracy and the Coda Lab accuracy. Coda lab accuracy is the accuracy we get when the model is tested on unseen test dataset experimented on submission to the EmoContext: Semeval Task 2019[1] .

From the Table 6.1, we conclude that the categorical scores and the dimensional scores boost the cross validation accuracy to 95% approximately and the test accuracy to 74% approximately. We also notice that the features obtained from the dependency parser which includes the subject and its actions, the objects and its actions do not have as significant result as compared to the emotional scores. Also categorical and emotional scores when taken along, play together a major boost in accuracy as compared to taking them individually. Thus, taking a cross model of dimensional and categorical emotion score model rather than the individual models and considering them as numeric variable

[1] https://www.humanizing-ai.com/emocontext.html

rather than a categorical variable played an important role in improving the performance of the emotion analysis.

Chapter 7

Goal 3: Implementation and Evaluation

7.1 Definition

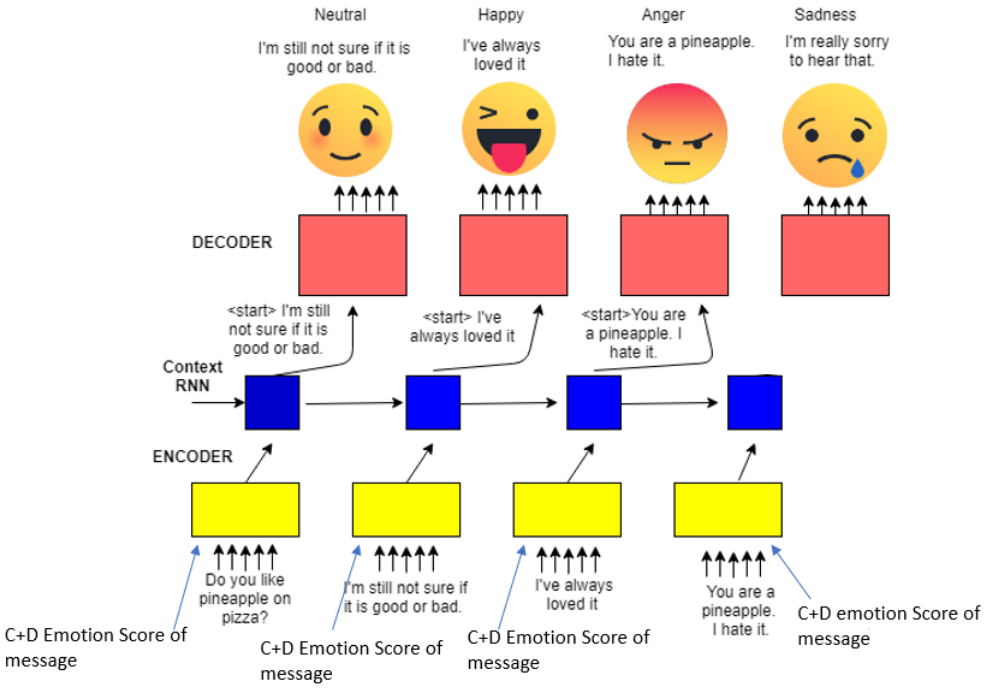
In this task, I used the developed corpus in Goal 1 and conclusions from the feature engineering experiments conducted in Goal 2 to finally feed the Hierarchical Recurrent Encoder Decoder architecture of the chatbot with the right set of features. The main task of this model is to generate emotional replies according to the emotion personality (Happy, Angry, Sad, Other/Neutral) given as an input by the user. A sample of the demo developed could be played at this https://harshitajhavar.wixsite.com/mysitelink.

7.2 Architecture

Team Replika [1] has released under Apache licence which allows their hierarchical recurrent encoder decoder architecture implementation source code for a chatbot generating replies based on categorical emotions to be reproduced and amended according to once own research requirements. I have improved on this for handling deep dialog context and for the Task 3, I also implemented my model of chatbot architecture as shown in Figure 7.1. Also, Replika team's data source is very small and is limited to only categorical emotions. While in my Master thesis, the data source is larger with 30k message conversations and I have worked robustly in Task 1 and Task 2 to make the corpus informed with both categorical and dimensional emotion model scores. Thus the feature glove vectors are also completely different. The thought vector is fed into the decoder on each decoding step. The decoder is conditioned on emotion labels. Both encoder and decoder contain 2 GRU layers with 512 hidden units each. The model is trained with

[1] https://github.com/lukalabs/cakechat

FIGURE 7.1: Architecture for Chatbot



context size 3 where the encoded sequence contains 30 tokens or less and the decoded sequence contains 32 tokens or less. It is initialized using message-emotion.scores pairs to vectors which are used as input to train on the corpus developed in goal 1 based on best features concluded in goal 2. We took only emotion scores for categorical and dimensional scores as these features gave the best results in Goal 2. The entire source code and related datasets is available at this [2] github repository.

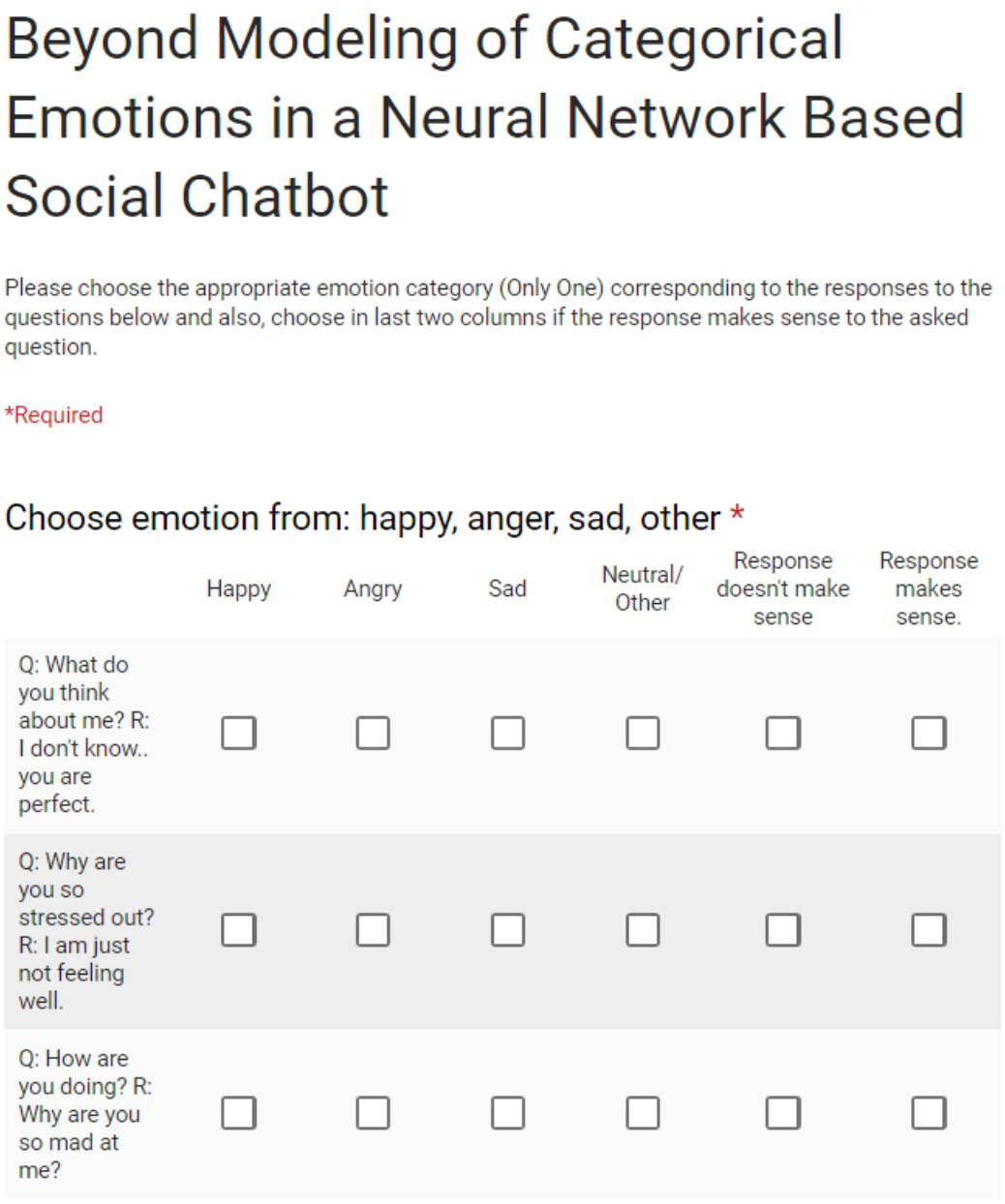
7.3 Evaluation

30 participants participated for performing human evaluation for this chatbot. The evaluation was conducted on Google Forms [3] . with query and corresponding reply by the chatbot as shown in the snapshot in Figure 7.2. There were total 30 queries printed alongside their replies.

Two things were tested from this human evaluation. One was identifying the emotion involved in the reply to a particular query and comparing it to the chosen input emotion label gold corpus while running the query with the chatbot. Second was if the response makes sense or doesn't make sense to evaluate the quality of the response with respect to the query.

[2] https://github.com/harshitaJhavar/ECB
[3] https://goo.gl/4cmsfw

FIGURE 7.2: Form for Human Evaluation: Task 3



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.

*Required

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?	☐	☐	☐	☐	☐	☐

7.3.1 Evaluating the emotion of the reply

The gold corpus to compare with what humans think of the emotion of the finally generated label as compared to the chosen emotion while running the query as input to a chatbot. The first four columns in Figure 7.2 were about this. The data collected from one participant is shown in Table 7.1 . The data analysis for the same participant is shown in Table 7.2. The conclusions for 30 participants is shown in Table 7.3.

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

TABLE 7.1: Data from one participant to compare the emotion of the reply with chosen input emotion

	TP	FP	TN	FN	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	
Sad	6	1	20	3	0.85	0.67	0.87	0.75
Other/Neutral	6	2	21	1	0.67	0.85	0.87	0.75

TABLE 7.2: Data Analysis on data collected from one participant to compare the emotion of the reply with chosen input emotion

	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/Neutral	0.752	0.791	0.853	0.7710

TABLE 7.3: Data Analysis From 30 participants to compare the emotion of the reply with chosen input emotion

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

TABLE 7.4: Data Collected from one participant to evaluate quality of response with respect to query

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

TABLE 7.5: Conclusion For 30 participants to evaluate quality of response with respect to query

7.3.2 Evaluating the quality of the response with respect to the query

The last two columns in Figure 7.2 were about this. The sample data analysis on performance from one participant is shown in Table7.4. The conclusions for 30 participants is shown in Table 7.5.

Thus, we conclude that more than 95% users found that the responses generated was sensible to the context of the question asked. Another conclusion from the human evaluation was 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%.

Chapter 8

Conclusion And Future Work

8.1 Conclusion

1. I developed a mapping between the dimensional and categorical model of emotions and thus, build a corpus labeled with the corresponding emotion scores for 30,160 message instances.
2. 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.
3. Performed emotion analysis of the user utterance .
4. I took as an input, the better emotionally informed corpus developed in goal 1 to the chatbot to generate emotional responses accordingly. With human evaluation, I concluded that more than 95% users found that the responses generated were sensible to the context of the question asked.
5. 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%.

8.2 Future Work

1. Improve the Feature Set: Semantic Role Modeling as another possible feature.
2. The current model could be further improved by making the chatbot to learn context personalization. For example: It will be interesting for the user to know not only the weather details but also recommendations for the activities which can be done by the user based on the past history of the user.

3. Domain understanding could be improved in a chatbot which can make a chatbot to incorporate more knowledge. For example: Doctor chatbots with lots of information about the medicine.

4. Abstraction can be performed in the quality of the response generated so as to have response in an understandable language to the common man rather than going into too technical details when asked about temperature, for example.

5. Develop better strategies to perform evaluation for chatbots. This is still state of art for research.

6. Extend the emotional conversational multimodal to speech output with emotional tones. A chatbot with emotional tones of surprise, sadness will be interesting to have and will make the user experience to the best.

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

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

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

Bibliography

[1] Matsuda. Frontiers in human neuroscience. 29(3):436–465, 2013.

[2] Heung-yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, Jan 2018. ISSN 2095-9230. doi: 10.1631/FITEE.1700826. URL `https://doi.org/10.1631/FITEE.1700826`.

[3] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR*, abs/1704.01074, 2018.

[4] Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *CoRR*, abs/1707.06996, 2017.

[5] Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/618. URL `https://doi.org/10.24963/ijcai.2018/618`.

[6] Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/P18-1104`.

[7] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995. Asian Federation of Natural Language Processing, 2017. URL `http://aclweb.org/anthology/I17-1099`.

[8] Saif M. Mohammad and Peter D. Turney. 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*, pages 26–34, 2010.

[9] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.

[10] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on 'Making Sense of Microposts'*, pages 93–98, 2011.

[11] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 01 2010.

[12] Saif Mohammad and Felipe Bravo-Marquez. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, 2017.

[13] Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In **SEM*, pages 65–77, August 2017.

[14] Harshita Jhavar and Paramita Mirza. EMOFIEL: mapping emotions of relationships in a story. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 243–246, 2018. doi: 10.1145/3184558.3186989. URL `http://doi.acm.org/10.1145/3184558.3186989`.

[15] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[16] Sven Buechel and Udo Hahn. Emotion analysis as a regression problem - dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI*, 2016.