Evaluating Significance of Emotion Classification in Emotion-Aware Empathetic Dialogue Systems

Akshen Kadakia akshenhe@usc.edu
Mrinal Kadam mrkadam@usc.edu
Swetha Sivakumar swethasi@usc.edu
Yashwanth Reddy Gondi gondi@usc.edu
Tanmey Saraiya tsaraiya@usc.edu

Abstract

Emotions are an important part of human communication. Distinguishing between them becomes very difficult because of the complexity and high interleaving tendency amongst these emotions. Most dialogue systems present today aren't good at judging the empathetic aspect of the conversation, leading to most responses being highly robotic and, in a few cases, even distasteful. In this project, our primary objective is geared towards identifying aspects that could help in achieving a more emotionally aware empathetic dialogue system to make the conversations flow more naturally and to help increase user engagement with the system. As a part of our experimentation, we are evaluating the impact of prepending emotion labels in empathetic dialogue systems. For this purpose we have experimented by adding different variations of emotion classifiers like LSTMs and Transformers on two different emotion datasets(8 classes and 32 classes). From the results obtained, we present the following observations: (1) Using a better emotion classification model to extend the retrieval model improves end-to-end performance (2) Making the emotion dataset less fine-grained with a lesser number of emotion classes improves the end-to-end performance. Though we cannot precisely quantify the observations due to the minimal number of training epochs, it can be clearly concluded that using a better emotion classifier model in the Emo-Prepend variant can improve the empathetic responding ability of dialogue systems.
The project video can be found at https://youtu.be/5JLwE3JxIaU.
The code of our project can be found at https://github.com/yashwanthreddyg/EmpatheticDialogues/.

1 Introduction

While it is straightforward for humans to recognize and acknowledge others' feelings in a conversation, this is a significant challenge for AI systems due to the paucity of suitable publicly available datasets for training and evaluation. A desirable trait in a human-facing dialogue agent is to appropriately respond to a conversation partner, by understanding and acknowledging any implied feelings - a skill referred to as empathetic responding. How to integrate empathetic responding into more general dialogue when, for example, the needs for empathy have to be balanced with staying on topic or providing information is also very important. These are the broad problems that we want to address.

Some examples of dialogues in the ED dataset grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding have been shown in Figure 1.

Our project aim is to evaluate a model's ability to produce empathetic responses and the impact of prepending emotion labels in empathetic dialogue systems. We are carrying out the following two experiments over the findings in the original base paper (Devlin et al. 2018) :

1. Changing the pre-trained emotion classifier used to train the 'Emotional Awareness' quotient of the dialogue system 2. Changing the EMPATHETICDIALOGUES (ED) dataset, a novel dataset with about 25k personal dialogues, given in the original paper, by reducing the number of categories of the target class(emotion) from 32 to a smaller subset of emotions, following which we will check for any significant increase/decrease in the retrieval model's performance.

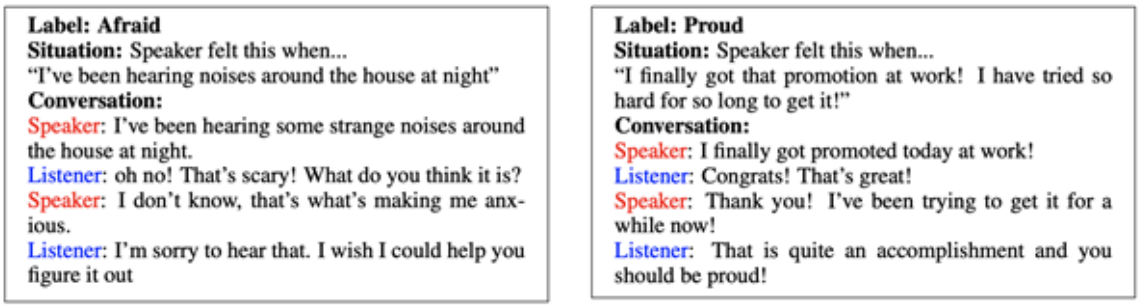

Figure 1 Examples of dialogues in the ED dataset. Taken from (Rashkin et al. 2019)

2 Methods

2.1 Task Formalization(Emotion Classsification):

Given a conversation sentence "x_1" as input, our emotion classifier model identifies an emotion label "y" for the same. We use 2 sets of "y":

1. original 32 emotions
2. 8 emotions derived manually from combining the 32 labels using references from Plutchik's Wheel of emotions

In order to keep the training times lesser while also accessing and utilizing extra data without modifying the base architecture, in the base paper, retrieval based empathetic dialogue generation model is extended with a supervised emotion classification model before feeding it into the encoder. This is referred to as the emo-prepend variation of the retrieval model.

In our project, we use three different external emotion classifiers to add supervised information to the data without making any modifications to the end-to-end model architecture. The emotion label that is predicted by the external classification model is then prepended to the input sentence which helps provide external supervised emotion information to the context and candidate encoder to retrieve an appropriate dialogue response to correspond with the user.

The average length of the sentences in the dataset is 17 words. Due to the textual nature and the length of the sentences in the dataset, we have based our models on LSTM and Transformer networks which have already been proven successful in the past for sentiment analysis tasks. To train the 'Emotional Awareness' quotient of the dialogue system, we use the following three models:

1. **BiLSTM —** These RNN models can model the temporal characteristics and the context of the data better by learning long term dependencies in the text. Originally proposed in (Hochreiter and Schmidhuber 1997)

2. **BiLSTM with Attention —** Attention can improve the performance of the vanilla LSTM model by extending the traditional encoder-decoder system by paying attention to specific words in the input sequence for each word in the output sequence. It's done by placing different focus on different words and by assigning each word with a score. Then using the softmax scores, the encoder hidden states are aggregated using their weighted sum to get the context vector. Originally proposed in (Vaswani et al. 2017)

3. **Transformer based model —** Transformers have proved to be even more effective than LSTMs as they are based on self-attention mechanisms. They are designed to take the whole input sequence at once and therefore significantly allow for more parallelization, thus reaching a new state of the art performance. Originally proposed in (Devlin et al. 2018)

3 Experiments

As a first step, we evaluate the ability of the supervised emotion classifiers used to predict the emotion to be prepended to the input. We then evaluate the models on their ability to reproduce the Listener's portion of the conversation (i.e. the ability to react to someone else's story).

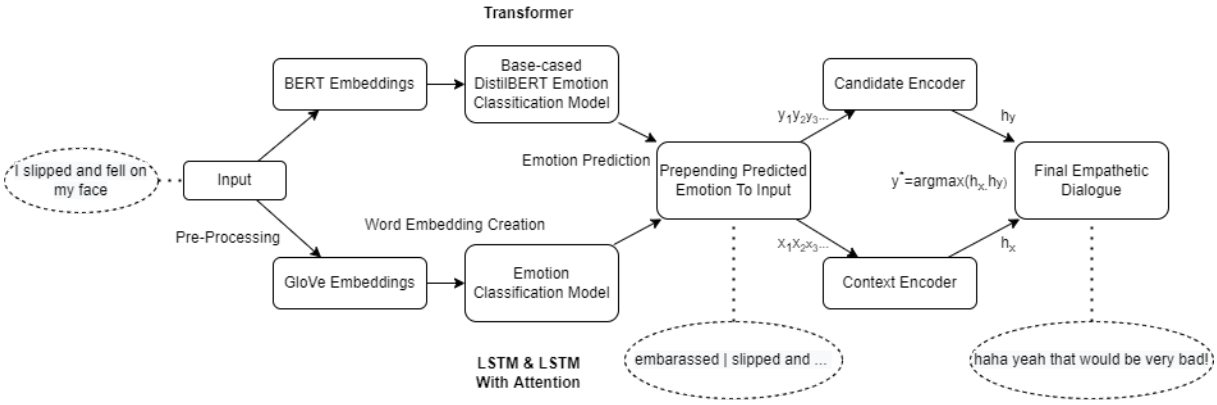

Figure 2 Architecture Block Diagram

3.1 Dataset

Dataset 1 (32 emotions): For our project, we use the publicly available EmpatheticDialogues Benchmark Dataset (Bouazizi and Ohtsuki 2019) with 25K conversations grounded in different emotions.

Dataset 2 (8 emotions): We group the 32 fine-grained target emotion classes into 8 broad categories according to Plutchik's wheel(Tromp and Pechenizkiy 2014) of emotions as seen in Figure 1. We do this because we achieve low accuracy as expected due to working on a relatively smaller dataset with such fine-grained emotion classes. We try to evaluate the notion that fine grained emotions might be hurting dialogue retrieval rather than helping, seeing the low accuracy of the classifier.

The distribution of the 8 broad categories after bucketing the original 32 emotions of the entire dataset are as shown in Table 1.

Emotion	Count	Distribution %
Sadness	29111	0.27
Joy	20885	0.19
Anger	14076	0.13
Fear	12796	0.12
Trust	11421	0.11
Surpise	8923	0.08
Anticipation	6621	0.06
Disgusted	3387	0.03

Table 1: Distribution of the 8 broad categories after bucketing the original 32 emotions of the entire dataset (Bouazizi and Ohtsuki 2019)

3.2 Empathetic Dialogue Systems(Multi-Class Emotion Classification) Algorithm

Algorithm 1 Multi-Class Emotion Classification

1: Build Vocabulary using ED Dataset
2: Add <UNK >and <PAD >to vocabulory
3: **if** Emotion_Classifer_model_type = "LSTM-Based" **then**
4: Build Embedding Matrix using Stanford's Pre-trained 100d GloVe Embedding
5: Label Encode the target emotion classes
6: **else if** Emotion_classifier_model_type = "Transformer-Based" **then**
7: Build Embedding Matrix using pre-trained BERT embeddings
8: **end if**
9: Split Train, Test and Validation set into features and target variables
10: **if** Train = True **then**
11: Train the classifier on Train Data
12: At every epoch, Save the best model till that point
13: If the validation accuracy doesn't change in over 10 epochs, Terminate Training
14: Run Accuracy Calculation Script
15: **else if** Test = True **then**
16: Load the best model
17: Predict emotions on Test Data
18: Write the emotion predictions to JSON file
19: **end if**
20: Prepend the predicted emotion to the input conversation and feed it to the Retrieval Model

Batch Size	64
Embedding Size	100
Max Sequence Length	512
Hidden Dimension	64
Number of Epochs	100
Seed	42
Optimizer	Adams
Loss Function	Categorical Cross-Entropy

Table 2: Hyperparameters used in the Emotion Classifier model

Model	8 Labels	32 Labels
Fast Text	-	21.21
BiLSTM	46.43	23.75
BiLSTM+ ATTN	**46.70**	**24.36**
DistilBERT	-	-

Table 3: **Emotion Classification Results.** Standalone emotion classification performance comparison. Best values are in bold.

3.3 Hyperparameters used in Model Training

For training Emotion classifiers, we use hyperparameters mentioned in Table 2. For training retieval with BERT models, we use the same hyperparmeters as used in (Rashkin et al. 2019).

4 Results and Discussion

Emotion Classification We calculate the accuracy score for all Emotion classifier models and report them in Table 3. We can see that `BiLSTM` models perform better than the `fasttext` model.

Retrieval w/ BERT We use the `BLEU` and `Precision` evaluation metrics as used in (Rashkin et al. 2019). The score for 32 emotion label set is reported in Table 4 and for 8 emotion label set in Table 5. The BiLSTM models perform better than the `fasttext` model overall. We also see improvements in the 8 emotion label dataset over the 32 emotion label dataset for the BiLSTM models.

5 Conclusion

Even though conclusive results cannot be stated by running the script for just 15 epochs, the initial results show a scope for improvement by using better emotion classification models and also by moving from a larger set of emotions to a smaller set. Thus, prepending the emotion label to the input helps in retrieving empathetic responses with the appropriate emotions.

References

Bouazizi, Mondher and Tomoaki Ohtsuki (2019). "Multi-class sentiment analysis on twitter: Classification performance and challenges". In: *Big Data Mining and Analytics* 2.3, pp. 181–194. DOI: 10.26599/BDMA.2019.9020002.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Rashkin, Hannah et al. (July 2019). *Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset*. Florence, Italy: Association for Computational Linguistics, pp. 5370–5381. DOI: 10.18653/v1/P19-1534. URL: https://aclanthology.org/P19-1534.

Tromp, Erik and Mykola Pechenizkiy (Dec. 2014). "Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel". In.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

	BLEU 1	**BLEU 2**	**BLEU 3**	**BLEU 4**	**Avg BLEU**	**P@1**	**P@1, 100**	**p@3, 100**	**P@10, 100**
FastText	14.10	5.16	2.55	1.42	5.81	53.03	65.08	83.60	95.63
BiLSTM	**14.16**	**5.20**	2.55	1.40	5.83	**53.05**	**65.17**	83.42	95.67
BiLSTM+ Attn	**14.16**	5.19	**2.57**	**1.43**	**5.84**	52.88	64.96	**83.44**	95.67

Table 4: **Performance Results for 32 Emotions(Multi-Class Emotion Classification).** End-to-End retrieval system performance comparison with 32 emotion classes. Best values are in bold.

	BLEU 1	**BLEU 2**	**BLEU 3**	**BLEU 4**	**Avg BLEU**	**P@1**	**P@1, 100**	**p@3, 100**	**P@10, 100**
BiLSTM	14.21	5.28	2.62	1.45	5.89	**53.30**	**65.04**	**83.52**	**95.62**
BiLSTM+ Attn	**14.24**	**5.30**	**2.66**	**1.49**	**5.93**	53.07	64.79	83.38	**95.62**

Table 5: **Performance Results for 8 Emotions(Multi-Class Emotion Classification).** End-to-End retrieval system performance comparison with 8 emotion classes. Best values are in bold.