SentEmojiBot: Empathising Conversations Generation with Emoji

Jatin Dholakia*, Amit Yadav*, Akhilesh Ravi*, Jainish Chauhan*, Naman Jain, Mayank Singh
Indian Institute of Technology Gandhinagar

Abstract

With increasing use of dialogue agent, it becomes extremely desirable that these agents understand and acknowledge the implied feelings and respond empathetically in any computer-mediated communication. Many traditional chatbots lack this crucial trait of expressing emotions. For humans, replying emphatically is straightforward. Emojis presents a promising way to express emotions using little faces. However, none of the AI systems uses emojis for empathetic conversation generation due to lack of any publicly available datasets and models to generate empathetic sentence containing emojis. This work proposes a novel dataset- SENTEMOJI that maps 25K empathetic conversations from EMPATHETICDIALOGUES dataset to emojis grounded on their emotions. Further, leveraging pre-exiting models and our dataset, we present an approach to generate empathetic sentences that include emojis without requiring lengthy full training of the model. Our user study indicates that dialogues from our model that use emojis outperforms same dialogues without emojis and were perceived to be more empathetic by human evaluators. Further, automatic metrics together with human evaluation indicate that both the dialogues with and without emojis generated by our model were relevant and empathising.

1 Introduction

Appropriately responding to a conversation is must by any human facing dialogue. Many studies (Reeves and Nass, 1996) show that humans interact with machines in a natural and social way, hence it reasonable to expect that the human facing dialogue agent also interact with them in similar way. Further many studies (Levinson et al. 2000, Wentzel 1997, Cassell 2001; Kim et al. 2004; Fraser et al., 2018) proved that social talk, acknowledging emotions, caring attitude and natural interaction, in fact, result in better task outcomes in diverse domains. Emojis are gaining popularity on social platform and personal chats to express emotions using little faces and humanise the text messages.

Figure 1: Empathetic Dialogues with emojis

There increasing use and humanise characteristics make them extremely promising to use in dialogue agent to improve their empathising trait. For instance, the dialogue from Figure1 (Rashkin et al., 2018) - “Congrats! That’s great!” is certainly empathising but by adding emoji at the end may make the dialogue more social, caring and engaging.

In this work, we investigate empathetic response generation with emojis. Training of language architecture on vast amount of social media conversations, books, and Empathetic dialogues (Ritter et al., 2010) (Zhang et al., 2018), (Torun, 1999)(Mazaré et al., 2019) have resulted in models that exhibit different behavioral responses including empathetic. Further, language model are trained to classify emojis (Çöltekin and Rama, 2018)(Barbieri et al., 2018)(Liu, 2018).

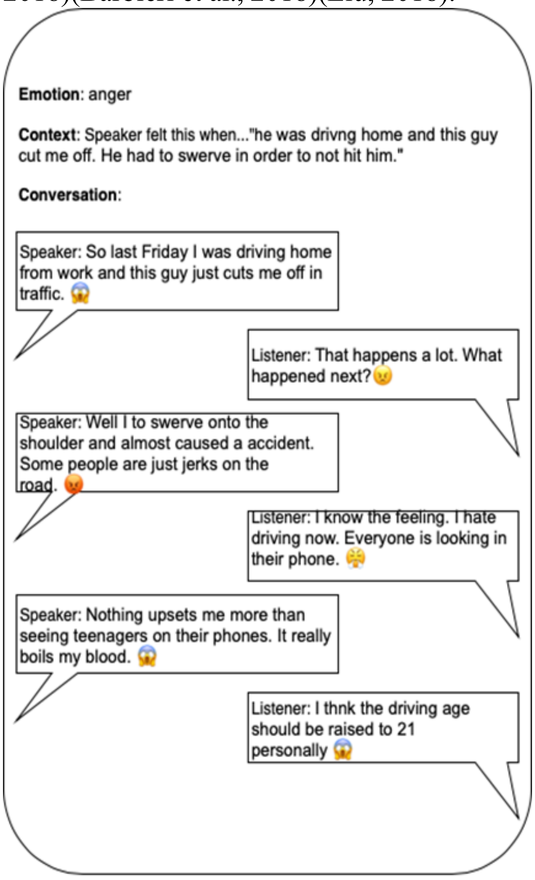

Figure 2: Example of conversation generation using our dataset. The Emotion label is shown from the actual ED dataset, the context is given to our dialogue agent. Clearly from conversation we can see that dialogue agent respond acknowledging the actual mapped emotion of the context with relevant emoji.

However, to the best of our knowledge there is no language model that generate empathetic conversation with emojis on ground of emotions of generated sentence.

This work aims to generate empathetic response with emojis. These emojis are included based in the emotion hidden in the generated sentence. We propose a new dataset SENTEMOJI (SE) that contains empathetic conversation from EMPATHETICDIALOGUE(ED) dataset, with each conversation based on their emotions mapped to an emoji. Figure2 presents an example of how ED dataset context is used to generate conversations with emojis. Clearly the generated sentence from the context understand emotions and generate empathising dialogues with emojis. The 32 emotions in ED dataset are reduced to 10 fundamental emotions (Kowalska and Wróbel, 2017), also this lead to combine the similar emotions. Moreover, this dataset is more balanced than the ED dataset. With this, a novel approach is presented to generate sentences with emojis by fine tuning a pre-trained BERT model. Further, a retrieval-based transformer model approach is also proposed. The code is available at https://github.com/JatinDholakia/SentEmojiBot .

Our experiments show that the dialogues generated with emojis are evaluated more empathising than the sentences without emojis. Moreover, the emojis and sentence generated were evaluated to be relevant to the conversation. The contribution of this work are thus: 1) we release a novel dataset that map 25K conversation from EMPATHETICDIALOGUE dataset with emojis; 2) we present a approach to generate empathising dialogue containing emojis; and 3) we showed that sentences with emojis are more empathising than sentence without emojis. Further paper is divided into sections: Relevant Work, Dataset, Methodology, Experiment and Evaluation. Finally, a Conclusion and Limitation of the current work is presented with possible future direction of the work.

2 Relevant Work

There has been a lot of work in the field of conversation generation(Zhou and Wang, 2018)(Zhou et al., 2018)(Wang and Wan, 2018)(Hu et al., 2017). Recently, there have been focus on generating empathising dialogue agents that understand and respond acknowledging the emotion of human (Rashkin et al., 2018)(Lin et al., 2019). Further, seeing increasing use of emojis researchers have worked on emoji prediction as well (Çöltekin and Rama, 2018)(Barbieri et al., 2018)(Liu, 2018). However, there have not been attempts to combine these two fields by generating empathetic responses containing emojis. Moreover, there is scope of improving the current maximum emoji prediction accuracy of 47% (Liu, 2018).

2.1 Empathetic Conversation Generation

(Rashkin et al., 2018) presents three models and a dataset ED for empathetic conversation generation. A transformer-based retrieval model, BERT-based retrieval model, BERT-based generative model has been proposed. In this method, these models have been pre-trained on large corpus, namely, 1.7 billion Reddit conversations, starting from scratch in case of the transformer model and from BERT base model released by (Devlin et al., 2018) for the BERT-based architectures. They were then fine-tuned for empathetic conversation generation on the ED dataset. Further, CAiRE (Lin et al., 2019) is an empathetic conversation generator adapted from the Generative Pre-trained Transformer (GPT) (Radford and Salimans, 2018). "CAiRE is built primarily to focus on empathy integration in fully data-driven generative dialogue systems." (Lin et al., 2019). Both of these work presents an approach to generate empathetic conversation but are limited in terms of using emojis to make conversation more engaging and empathising.

Basic Emotions	Emotions in ED
Fear	Afraid, Terrified, Anxious, Apprehensive
Anger	Angry, Furious, Annoyed
Disgust	Ashamed, Guilty, Embarrassed, Disgusted
Sadness	Sad, Sentimental, Nostalgic, Disappointed
Contempt	Lonely, Devastated, Jealous
Amusement	Surprise, Excited, Anticipating
Pride in achievement	Proud, Impressed
Satisfaction	Grateful, Prepared, Content
Optimism	Confident, Faithful, Trusting, Hopeful
Contentment	Joyful, Caring

Table 1: Mapping of ED dataset to SE dataset

2.2 Emoji Prediction

"SemEval 2018 Task 2: Multilingual Emoji Prediction" (Çöltekin and Rama, 2018)(Barbieri et al., 2018)(Liu, 2018) was organised as a part of SemEval 2018. In this task, the text of tweets was given, and the emoji to be used is to be predicted for each tweet. (Çöltekin and Rama, 2018) have shown that the proposed support vector machine (SVM) model performs better than the recurrent neural networks (RNNs). The SVM uses bag-of-word/character n-gram features as input. The proposed RNN is a bidirectional RNN that uses word and character sequences as input. This was the work that performed the best in the task with a macro F1-score of 36% and accuracy of 47%. (Liu, 2018) has proposed a Gradient Boosting Regression Tree Method and a Bidirectional Long-Short Term Memory Network model for this task. This also gives a performance close to (Çöltekin and Rama, 2018). However, there is a lot of scope of improvement in this area. And further in lack of any emoji mapped empathetic conversation dataset these works are limited in terms of using emoji classification for tasks like generating emojis for emotional empathetic conversations. We have built on the current literature to propose a method combining empathetic conversation generation and emoji prediction to generate empathising conversations with emojis.

3 Dataset

This section details the ED dataset used for preparing SENTEMOJI(SE) dataset. Further it explains mapping of emojis and preparation of SE dataset. At last of this section, an analysis of SE dataset is given.

3.1 EMPATHETICDIALOGUE DATASET

The ED dataset consists of 24,850 conversations. It consists of a total of 79,190 utterances with each utterance of average length equal to 15.2 words. Along with each conversation its context and its emotion are also given in the dataset. Overall conversations were classified in 32 emotions with average share of 3.125%.

3.2 SENTEMOJI DATASET

The empathy dialogues dataset had utterances and their emotions but it didn't have any emojis. Also, it consisted 32 emotions, in which very less frequently used emotions were also included and many of the emotions were very overlapping. According to a book by (Kowalska and Wróbel, 2017) on Basic Emotions there are 10 basic emotions. Emotions which differ from each other in terms of their appraisal, antecedent events, probable behavioral response and physiology are classified basic emotions. According to this there are 5 negative emotions: and 5 positive emotions. So, for SE we have remapped 32 emotions in the ED dataset to the 10 basic emotions. Table 1 show mapping of ED dataset to SE.

Our SE dataset has 10 columns. The last two columns contain emojis corresponding to each utterance and UNICODE of each emoji respectively. The other columns have been derived from the ED dataset (Rashkin et al., 2018). So, overall the SE dataset has the following columns: conv_id, utterance_idx, context, prompt, speaker_idx, utterance, selfeval, tags, emoji, emoji_unicode.

For each of the 10 basic emotions, a few of the most commonly used emojis have manually been added by us. The most commonly used emojis

Emotion	Emojis
Fear	😳, 😰, 😨, 😬
Anger	😱, 😠, 😤, 😒, 😑, 😠
Disgust	😭, 😩, 🤦‍♀️, 😯, 🤢, 😖
Sadness	😫, 😭, 😞
Contempt	😔, 💔, 😏
Amusement	😲
Pride in achievement	👌, 👏, 👍
Satisfaction	😊, 🙏, 😔, 😎
Optimism	💪, 😎, 🛐
Contentment	😇, 😘, 😂, 😁, 💃

Table 2: Emoji bucket for each emotion

were obtained from (Novak et al., 2015). The list of emojis for the 10 basic emotions (Kowalska and Wróbel, 2017) are shown in Table 2. Figure 3 shows the distribution of utterances in the new dataset. The average share of each emotion is 9.99% and the standard deviation of 2.28%. Clearly, the distribution is not biased much towards any of the emotion

For adding emojis to the utterances, the utterances with pre-labelled emotions have been taken. Then, for each utterance, an emoji is selected from the bucket of emojis for that emotion. This is done using Word2Vec and Emoji2Vec models. The utterance is converted into a vector using Word2Vec and the vector representation of all the emojis in the bucket are obtained using Emoji2Vec.

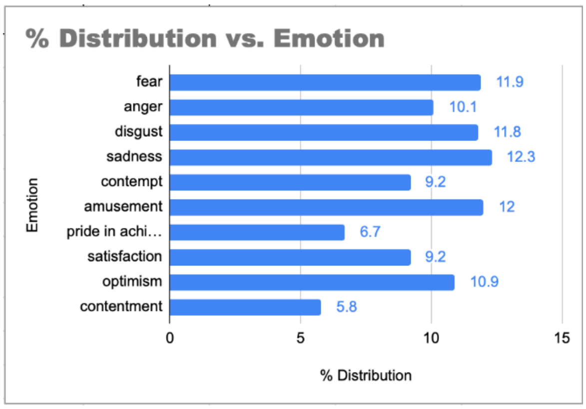

Figure 3: Distribution of conversations labels within SENTEMOJI dataset

The vector obtained for the utterance is compared with the vectors of all emojis using cosine similarity. Based on this, the closest emoji is selected and, finally, this emoji is added to the utterance. More details about Word2Vec and Emoji2Vec are given in the Methodology section.

4 Methodology

In this section, we discuss how we have used various architectures to generate conversation responses in the *Listener* role. In order for our models to emulate real life conversations, it is necessary that the models have only the information that were present with the speaker and listener. Hence the models were only given the previous utterances from the conversation - including dialogues from both the *Speaker* and *Listener*.

We divide our approach into two sub-tasks - generating response sentences and adding relevant emoji to the sentences. For generating a response, we have experimented with a Transformer based retrieval architecture and BERT based model. For adding emojis to sentences, we have used emoji2vec, word2vec and a CNN based emotion classifier. These approaches are explained in detail below.

4.1 Preprocessing

The utterances were separated into context and response utterances based on the speaker_id. All previous utterances in a conversation are concatenated to form context and listener's utterance forms the response. To tokenize the utterances we have used BertTokenizer(Thomas et-al.,2019) which uses WordPiece embeddings with a 30,000 tokens vocabulary. The sentences are concatenated to form a sequence with a separator token "[SEP]" between them. The tokens are converted to vectors according to their indices in the dictionary. The sequences are made to have a maximum length of 100. Sequences with larger lengths are trimmed to take last 100 tokens. Sequences with lengths smaller than 100 are padded to make their length equal to 100.

4.2 Retrieval Based Architecture

In retrieval-based architecture, the model is given a set of candidates Y, from which the model chooses the *"best"* candidate according to the softmax of the dot product of the encoded context and candidates. The candidate set Y consists of all the responses in the batch. For simplicity let us call the context sequence x and the response sequence y both of which are vectors of dimension 100 each. The encoded context sequence h_x and the encoded response sequence h_y.

The context and set of responses are encoded using a transformer networks (Vaswani et at., 2017) which have proven success in dialogue generation (Zhang et al., 2018). Separate encoders are used for context and response encoding. The model is trained to maximize the likelihood of producing the correct target response given context, i.e. p(ȳ|x), where ȳ is the predicted response from the SoftMax layer and x is the context sequence. To achieve this the negative log likelihood is taken as error and minimised. The first

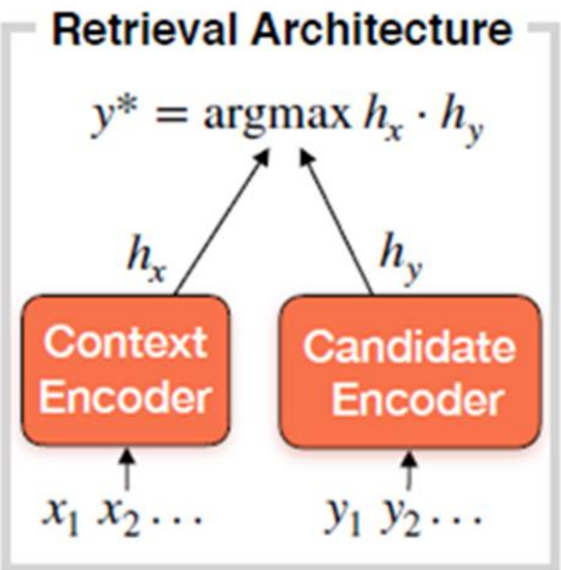

Figure 4: Retrieval-based architecture as proposed by (Rashkin et al., 2018)

layer of the encoder is a 300-d embedding layer. It is followed by 4 multi-headed attention layers. It is followed by a LayerNorm layer, a dropout layer and another LayerNorm layer. The model has close to 16M parameters.

4.3 Fine-tuning BERT

We also experimented with pretrained BERT[1] based model which provides a general purpose architecture for Natural Language Understanding and has been successful for the next sentence prediction. We use the *'bert-base-uncased'* pretrained model by HuggingFace (Thomas et-al.,2019).

BERT was trained to perform well on fine-tuning tasks which was the reason we chose it for our purpose. Pre-training of $\mathrm{BERT}_{\mathrm{base}}$ was performed on BookCorpus (800M words) and English Wikipedia (2500M words). To train BERT a masked language model (MLM) was used which randomly masks some percentage of the input and the aim is to identify the index in vocabulary of the word. This is different from the transformer model which uses a left-to-right language model. $\mathrm{BERT}_{\mathrm{base}}$ consists of 12 layers of transformer blocks and it uses an embedding size of dimension 768. The number of self-attention layers are equal to 12. The total number of parameters are close to 110M.

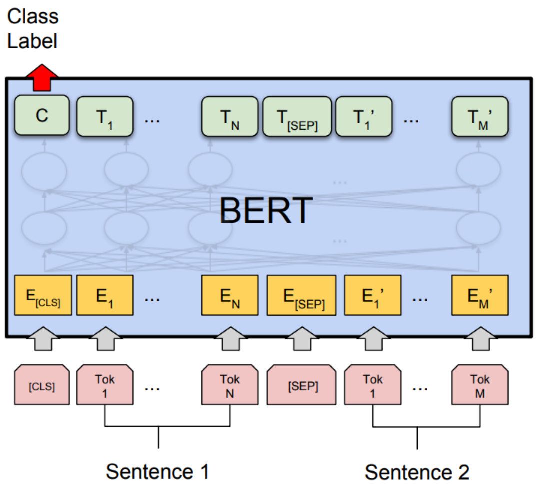

Figure 5: Overall architecture of BERT showing input output format as proposed by (Rashkin et al., 2018)

4.4 Adding Emoticons to Sentences

This stage involves two steps: firstly, the identification of the emotion of the utterance and, secondly, obtaining the appropriate emoji for the utterance from the list of emojis corresponding to that emotion.

For **CNN based emotion classifier**, **pre-processing** was performed. The non-alphabetic characters are removed from each utterance. Then, the Keras tokenizer, with the nb_words (number of words) set as 20000, was fit on the train data. Then, each utterance was converted to a sequence, padded and converted to a 1000-dimension vector using the above obtained tokenizer to generate vectors for each train and test utterance.

The **CNN** architecture has been adapted from the convolutional neural network proposed by (Kim, 2014) for sentence classification. It consists of four layers (with dropout): 1) An embedding layer with input dimension of 1000; 2) A 1-D convolutional layer 300 filter each of length 3 and A tanh activation is used; 3) A fully-connected

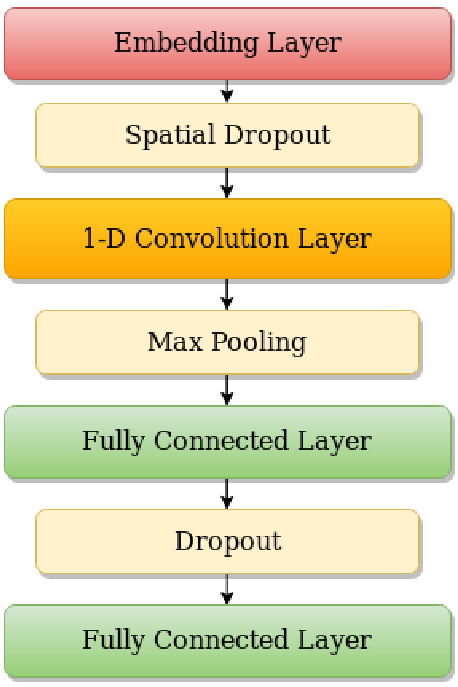

Figure 6: Architecture of CNN-based Emotion Classifier

dense layer of dimension 300 with a ReLU activation is used; and lastly 4) A fully-connected dense layer of dimension 10 (number of classes is 10). With a sigmoid activation is used.

Finally, after getting emotion of generated sentence a emoji was selected from the mapped bucket of emoji using **Word2Vec** (Demeester et al., 2016) **and Emoji2Vec.**(Eisner et al., 2016)

The pre-trained **Word2Vec** (Demeester et al., 2016) model is used to obtain the word vector for the words present in the utterances. The Word2Vec model was pre-trained by Google on a part of "Google news" dataset which contains 100 billion words. The model consists of 3 million words and phrases. The word vector of each word obtained from Word2Vec is of dimension 300.

The pre-trained **Emoji2Vec** model (Eisner et al., 2016)is used for obtaining the vector representation for each emoji. The vector for each emoji is of dimension 300 and Emoji2Vec is pre-trained in such a way that the vector for each emoji is close to the vectors for words which have similar meanings to the corresponding emoji. The model was pretrained on a twitter dataset. The dataset used for pretraining consists 67000 English tweets. Each tweet was labelled as positive, neutral or negative manually. 46% tweets are labelled positive, 29% tweets are labelled neutral and 25% tweets are labeled negative.

After the response is generated, the above classifier is used to obtain the corresponding emotion for the response. Then, word vectors are obtained for all the words in the response using Word2Vec. This is compared to the Emoji2Vec vectors for all the emotions pertaining to the emotion. The closest emoji is selected by using cosine similarity. This emoji is added to the generated response.

5 Experiment and Evaluation

In this section we will discuss about our experiments details to evaluate the generated dialogues. Further it will provide the results of automated metrics and human evaluation.

5.1 Experimental Setup

Retrieval-based transformer model, in this approach, the transformer architecture shown in Figure 4. is used. The utterances from the ED dataset have been split into three parts and used for training data, validation data and test data for the transformer with a split of 80%, 10% and 10% respectively. As mentioned earlier, the utterances were separated into context and response utterances based on the speaker_id and preprocessing is done as stated earlier. In the training phase, the learning rate was set to 8 x 10^{-4} and the optimizer choice was Adamax. The model was trained for 25 epochs with a batch size was set of 128.

BERT-based model, in this approach, the BERT architecture shown above is used. The train, validation and test data and the split are same as that used for the retrieval-based transformer model. As mentioned earlier, the utterances were separated into context and response utterances based on the speaker_id and preprocessing is done as stated earlier. Since $BERT_{base}$ has already been pre-trained, we directly went ahead with the fine-tuning phase. A BERT embedding dimension of 300 is used. The learning rate was set to 5×10^{-5} and the optimizer choice was Adamax. The model was fine-tuned for 12 epochs with a batch size was set of 16.

CNN-based Emotion Classifier was trained on the original contexts, i.e., the first utterance for each conversation in the ED dataset. Each context has an emotion label associated with it in the ED dataset (32 emotion labels). These labels are replaced with the corresponding label from the set of 10 emotion labels mentioned earlier in the "Datasets" section as shown in table1. In the ED dataset, there are 20628 original contexts. This set of utterances (original contexts) is split into 80% for train data and 20% for test data. The train data was further split into 90%-10% for a new train data and validation data. This new train and validation were used for obtaining the best hyperparameters. Once these were obtained, the original train data was used to train the model. An Adam optimizer is used with a learning rate of 0.001 and a decay of 10^{-6}. This model is trained for two epochs with a batch size of 128 (experimentally, on training for more than two epochs we found that the model was overfitting on the train data).

Model	**BLEU**	**P@1,100**
Transformer	4.38	3.65%
BERT	5.78	36%

Table 3: Automated evaluation metrics for sentence generation on the test set.BLEU and P@1,100 scores for the Transformer model and BERT.

5.2 Evaluation

For **Automated Evaluation Metrics,** The average BLUE score (average of BLEU-1, BLEU-2, BLEU-3 and BLEU-4) is computed for the retrieval-based transformer model and BERT-based model. Along with this, p1@100 is also computed and used for evaluation. p1@100 is the accuracy of the model at choosing the correct response out of a hundred randomly selected examples in the test set.[2] For the evaluation of the CNN-based emotion classifier, macro F1-score and macro accuracy are used.

For **Human Evaluation**, we conducted a user study for measuring the empathy of our sentences and the relevance of the emojis that we have added. We took 50 sentences given as output by our model which had emojis. And we took 50 sentences given as output by our model which didn't have emojis. We provided human evaluators with these output sentences and the inputs given to the model for which these outputs were generated. We asked them to give a score to the empathy of the outputs and the relevance of the emoji on a scale of 1-5. Each user was randomly given 20 sentences for evaluation. Each sentence was given a score by 5 users. The scores obtained are first averaged over the users and later over all the sentences.

User-Study	**Average Empathising Score (1-5)**	**Relevance of emoji (1-5)**
Responses without emojis	2.9 / 5	NA
Responses with emojis	**3.3 / 5**	3.03 / 5

Table 4: Human evaluated metrics. The average score of responses generated with and without emojis and score of relevance of emoji added

6 Inferences and Conclusion

BERT based encoder performed better than Transformer model for response generation. The BERT base model has already been pre-trained by Google, only fine-tuning is required for obtaining good performance metrics. It gives performance comparable to that given by the BERT base which was both pre-trained and fine-tuned by (Rashkin et al., 2018). The results showed that the sentences with emojis were more empathising than the sentences without emojis. Also, the score of relevance of emojis was also good which implies that the CNN based emotion classifier was able to find the emotion correctly and the emojis assigned were relevant.

Hence, as the empathising scores for sentences with emoticons were better than those without. This indicate that emojis can empathise more. And hence, in future it will be worth trying other language models to generate empathising sentences with emojis.

7 References

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 Task 2: Multilingual Emoji Prediction. :24–33.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in Emoji Prediction. :34–38.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*:1389–1399.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Mlm).

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. :48–54.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. *34th International Conference on Machine Learning, ICML 2017*, 4:2503–2513.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*:1746–1751.

Magda Kowalska and Monika Wróbel. 2017. Basic Emotions.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2019. CAiRE: An End-to-End Empathetic Chatbot. :1–2.

Man Liu. 2018. EmoNLP at SemEval-2018 Task 2: English Emoji Prediction with Gradient Boosting Regression Tree Method and Bidirectional LSTM. :390–394.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2019. Training Millions of Personalized Dialogue Agents. :2775–2779.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE*, 10(12):1–22.

Alec Radford and Tim Salimans. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI*:1–12.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset.

Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, New York, NY, US.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*(June):172–180.

Erdal Torun. 1999. UAV Requirements and Design Considerations. *RTO SCI Symposium on "Warefare Automation: Procedures and Techniques for Unamanned Vehicles"*:B4-1-B4-9.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. *IJCAI International Joint Conference on Artificial Intelligence*, 2018-July:4446–4452.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2204–2213.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*:730–738.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. *ACL 2018 - 56th Annual Meeting of the*

Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1:1128–1137.

BERT — transformers 2.1.1 documentation.