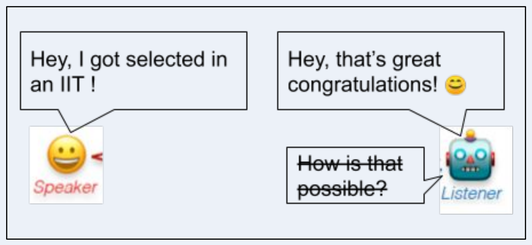
SentEmojiBot : Empathising Conversations Generator

Chauhan Jainish, Jatin Dholakia, Akhilesh Ravi, Amit Yadav

Mentor: Prof. Mayank Singh, Naman Jain

Problem Statement and Motivation

- Traditional chatbots are unable to recognise the feelings of the person and reply accordingly.
- They also lack the ability to use emojis effectively for creating an engaging and empathising conversation.
- Emojis are gaining popularity in day-to-day computer-mediated conversations (CMCs), resulting in more interactive conversations.
- Prior work has either classified the emojis or generated empathetic dialogue without the use of emojis.
- The use of empathising replies in CMCs makes it more relatable and emojis improve it further. When a chatbot uses emojis, it can become more like a human in conversation.



Previous work

- **Empathetic Dialogues (Facebook Research):** This proposed a new benchmark for empathetic dialogue generation. It has three approaches - using transformers and BERT to create an empathetic dialogue agent. It also proposed a dataset, Empathetic Dialogues.
- **CAiRE:** An End-to-End Empathetic Chatbot - It is an empathising chatbot that takes into account the persona and a history of a few utterances for generating a response.
- **SemEval 2018** Tasks: Emoji Prediction

Datasets

Empathetic Dialogues dataset:
- Columns: conversation id, utterance id, context, prompt, speaker id, utterance, self evaluation
- 24,850 conversations, 79190 sentences, 32 emotions

Emoji2vec dataset:
- Trained from Unicode descriptions.
- Trained over 67k English tweets labelled manually for positive, neutral or negative sentiment.

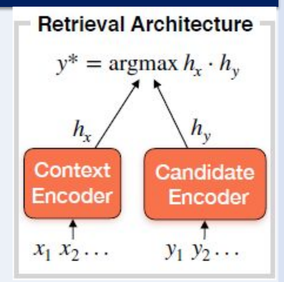
Mapping of Emotions and Emojis:
- Manually annotated most frequently used emojis into 10 emotion classes.

SentEmoji dataset:
- Dataset generated by us using Empathetic Dialogues dataset, emoji2vec model and word2vec model. (example of conversation in figure).
- In each utterance, emoji has been added and also the number basic emotions are taken to be 10
- Dataset accepted in CoDS-COMAD 2020

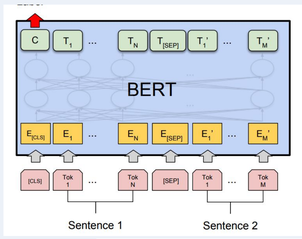
Implementation

Preprocessing
- Tokenizing utterances using Bert Tokenizer.
- Adding special tokens like [CLS] and [SEP], to mark start and end of utterance.
- Converting to vectors using their indices from dictionary.
- Padding the sequences to make them of same length as others in batch.

Retrieval based architecture
- Two encoders separately encoding the context and target.
- Candidates are formed by encoding all target sequences in batch.
- Model chooses the candidate utterance which maximises the softmax dot product $h_x.h_y$.

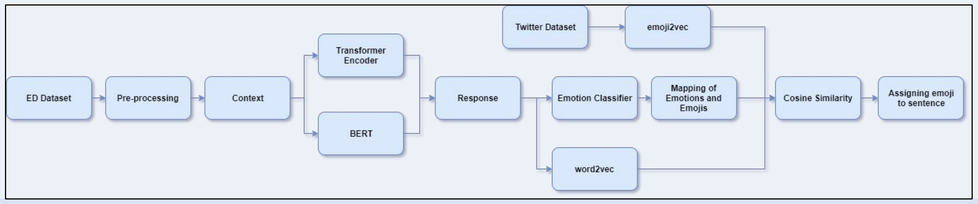


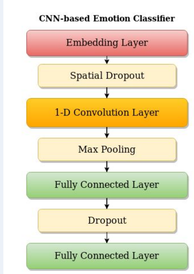
Fine-tuning BERT
- Used Pre-trained BERT implementation provided by Hugging Face.
- Fine-tuned on ED dataset.
- Retrieval based architecture.



CNN based emotion classifier
- Takes a sentence/utterance.
- Tokenizes it and converts it into a vector of 1000 dimensions
- Trained on the contexts of the the Empathetic Dialogues dataset for 10 emotion classes

Emoji2vec
- Pre-trained on twitter dataset using skip-gram method.





Results

- BERT performs better than the transformer for conversation generation.
- Emojis have an positive impact in making the conversation empathetic.

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

User Study	Average Empathising score (1-5 scale)
Responses without emojis	2.9 / 5
Responses with emojis	**3.3 / 5**

Emotion Classifier

Relevance of emoji (user study) (1-5 scale)	3.03 / 5
Macro F1 score	56%
Macro-accuracy	58%

References

- Ben Eisner, Isabelle Augenstein, Tim Rockta ̈schel, Matko Bos ̃njak, Sebastian Riedel emoji2vec: Learning Emoji Representations from their Description *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, November 1, 2016.
- Chatbot HTML template : https://github.com/sahil-rajput/Candice-YourPersonalChatBot
- Mikolov, Tomas & Sutskever, Ilya & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems. 26.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, Y-Lan Boureau, Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset
- Frequently used emojis: https://www.kaggle.com/thomasseleck/emoji-sentiment-data
- Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and R'emi Louf and Morgan Funtowicz and Jamie Brew, HuggingFace's Transformers: State-of-the-art Natural Language Processing
- https://gluon-nlp.mxnet.io/examples/sentence_embedding/bert.html
- ES 654 Project by Balani Mohit, Kaushal Modi and Akhilesh Ravi