Empathic Response Generation in Chatbots

Timo Spring
University of Bern
Bern, Switzerland
timo.spring@students.unibe.ch

Jacky Casas
HES-SO
University of Applied Sciences
Western Switzerland,
Fribourg, Switzerland
jacky.casas@hes-so.ch

Karl Daher
HES-SO
University of Applied Sciences
Western Switzerland,
Fribourg, Switzerland
karl.daher@hes-so.ch

Elena Mugellini
HES-SO
University of Applied Sciences
Western Switzerland
Fribourg, Switzerland
elena.mugellini@hes-so.ch

Omar Abou Khaled
HES-SO
University of Applied Sciences
Western Switzerland
Fribourg, Switzerland
omar.aboukhaled@hes-so.ch

Abstract

Recent years show an increasing popularity of chatbots, with latest efforts aiming to make them more empathic and human-like, finding application for example in customer service or in treating mental illnesses. Thereby, emphatic chatbots can understand the user's emotional state and respond to it on an appropriate emotional level. This survey provides an overview of existing approaches used for emotion detection and empathic response generation. These approaches raise at least one of the following profound challenges: the *lack of quality training data*, *balancing emotion and content level information*, considering the *full end-to-end experience* and *modelling emotions throughout conversations*. Furthermore, only few approaches actually cover response generation. We state that these approaches are not yet empathic in that they either mirror the user's emotional state or leave it up to the user to decide the emotion category of the response. Empathic response generation should select appropriate emotional responses more dynamically and express them accordingly, for example using emojis.

1 Introduction

Chatbots are everywhere, from booking a flight online to checking the balance of a bank account. Thereby, most of these interactions with chatbots are still of transactional nature, for example when ordering a pizza. Furthermore, the interactions with chatbots are usually short and therefore, not resembling normal human-like conversations. Hence, recent efforts aim to also create more personalised chatbots for deeper and emotionally charged conversations. This can also help boosting the usage of chatbots by making users feel better, instead of providing or offering certain services to them. Thus, making the overall interaction more natural and human-like. Empathic chatbots find application for example in customer service or for treating mental illnesses.

Chatbots for customer service is a growing trend and *Gartner*[1] predicts that by 2020 about 25% of customer service requests will be handled using chatbots. Xu et al. (2017) have analysed one million service requests made over Twitter. The authors note that about 40% of the requests express emotions, attitudes or opinions rather than seek for specific information. In addition, the average response time for customer service requests is about 6.5 hours. However, 72% of users who file a request, expect a response within an hour. Thus, empathic chatbots could help improving customer support by reducing the response time, reacting to specific user emotions and reduce overall costs.

Empathic chatbots also indicate potential in di-

[1] https://www.gartner.com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtual-customer-assistants-by-2020

agnosing and treating mental illnesses. According to the Swiss Health Observatory, one out of five Swiss suffers from at least a slight depression[2]. In the United States of America, nearly one in five adults suffers from some form of mental illness causing economic costs of around $210 billion annually[3]. A lack of mental professionals and psychiatrists makes it difficult to treat and detect affected individuals. Empathic chatbots can provide good accessibility and are scalable to a vast public with a low entrance-barrier and to help detecting and treating mental illness faster.

There are already some noteworthy advancements in the field of empathic chatbots to treat or detect mental illnesses. *Woebot* (Fitzpatrick et al., 2017) for example, is a chatbot from the University of Stanford, using methods from Cognitive Behavioural Therapy (CBT) to provide a step-by-step guidance to users with anxieties or depressions. Another noteworthy chatbot is *Replika* in the form of a digital companion with the main goal of providing someone to talk to 24/7 and to tackle certain resolutions for example being more social[4]. The underlying code of *Replika* is open source.

One important aspect in designing empathic chatbots is understanding what empathy actually is. In this survey, we consider empathic chatbots to use affective empathy as defined by Liu and Sundar (2018). So, the chatbots detect and understand the user's emotions and respond to them on an appropriate emotional level. Liu and Sundar (2018) observe in their study that the expression of either sympathy or empathy from a health advice chatbot is favoured over an unemotional response.

With chatbots becoming more and more human-like, it gets difficult for people to distinguish online conversations with bots and humans. This fact has lately become problematic[5], since bots are increasingly being misused for political propaganda and the manipulation of people. The Computational Propaganda Research Project (COMPROP)[6] from the University of Oxford devotes it's time to investigating, how chatbots and other algorithms are used to manipulate the public and form opinions, yielding in multiple reports on the matter. Woolley and Guilbeault (2017) analysed the usage of chatbots during the 2016 presidential election in the United States using a quantitative network analysis of over 17 million tweets. The authors state that chatbots in fact showed a measurable influence during the election by either *manufacturing online popularity* or by *democratizing propaganda*. Thus, governments of several countries start to introduce regulations to fight against these kinds of online manipulations (Howard et al., 2018). However, chatbots oftentimes remain a widely-accepted tool for propaganda (Woolley and Guilbeault, 2017).

The rest of this survey is structured as follows. Section 2 introduces the different stages in the interaction with empathic chatbots. For each stage, we present the most common and noteworthy approaches. In Section 2.3, we focus on the state of empathic response generation and outline shortcomings. Finally, we conclude and discuss the survey in Section 3.

2 The Four Stages of Empathic Chatbots

We partition the interaction with an empathic chatbot in four stages — the *emotion expression* by the user in text format, *the emotion detection* and *response generation* by the chatbot and the *response* or rather *emotion expression* from the chatbot back to the user in text format. An overview of the stages can be seen in Figure 1. Each stage requires special attention to ensure a proper end-to-end user experience. The following chapter presents each stage and points out common approaches, challenges and shortcomings.

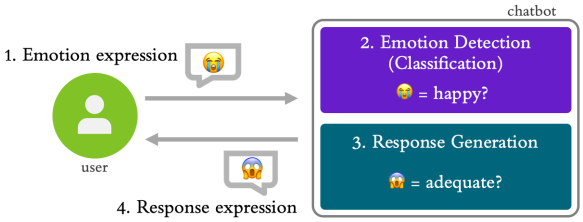

Figure 1: Four stages of interaction to consider when building empathic chatbots

2.1 Emotion Expression

Emotions are a complex construct and the ability to detect emotions in text is heavily dependent

[2] https://www.obsan.admin.ch/de/publikationen/psychische-gesundheit

[3] https://www.nimh.nih.gov/health/statistics/mental-illness.shtml

[4] https://replika.ai

[5] https://www.nytimes.com/2018/12/04/opinion/chatbots-ai-democracy-free-speech.html

[6] https://comprop.oii.ox.ac.uk

on how these emotions are expressed. Even for humans, it can be tricky to guess the emotional state of a text message. There are three major challenges when it comes to emotions. First, they are *context sensitive* by nature, they can be *multi-layered* within a sentence, and they can be *implicit*. Hence, emotions are perceived differently based on contextual and personal circumstances, such as the culture, age, sex, education, previous experiences and other individual parameters (Ben-Zeev, 2000; Oatley et al., 2006).

In normal face-to-face conversations, emotions are also expressed over the tonality of the speaker, body language, gestures and facial expressions. However, when focusing solely on the emotion expression in text, lots of potential information stemming from these non-verbal cues go lost. This might lead to mis-interpretations of the opponent's emotions, when communicating over text messages only.

Words holding a strong emotional charge such as *kisses* for *love*, *tears* for *sadness*, or *wow* for *surprise* can help interpreting the emotional state. Such word associations are also used in *emotion lexicons* and *word embeddings*.

The usage of emojis can help amplifying or transporting emotional meaning in text-based conversations, but can also pose additional interpretation challenges, for example, when multiple contradicting emojis are used. We further discuss emojis in the context of response expression in Section 2.4.

2.2 Emotion Detection

In the emotion detection stage, we try to classify and map an utterance to an emotional category. It is important to note, that emotion detection is strongly tied with response generation as similar approaches are used for both stages.

One of the first challenges is setting the number of emotion categories to be used for classification. There is no universally accepted model of emotions and the number of emotions differs drastically depending on the underlying model. One of the most popular models in emotion detection is *Ekman's six basic emotions — happiness, sadness, fear, anger, disgust,* and *surprise* (Ekman, 1992). Other popular models include *Plutchik's wheel of emotions* (Plutchik, 1991) or *Parrot's Emotional Layers* (Parrott, 2001) consisting of thirty-one different emotions. However, the latter two models are seldomly used for emotion detection, since more emotion categories mean additional complexity for the classification task.

Seyeditabari et al. (2018) review existing works and approaches in the field of emotion detection and provide a good overview of the state-of-the-art of emotion detection in text. They list different resources used for detecting emotions in text such as labelled text, emotion lexicons, or word embeddings and elaborate on common approaches used for emotion detection. Seyeditabari et al. (2018) conclude that there is still potential for improving emotion detection in text. Thereby, the complex nature of emotion expression, the shortage of quality data and inefficient models induce most challenges for future work.

In this survey, we distinguish three major approaches used for emotion detection in text — *rule-based, non-neural machine learning* and *deep learning*.

2.2.1 Rule-Based Approaches

Rule-based approaches mainly use *emotion lexicons* or *word embeddings*. Both approaches are based on keyword lookup from text to detect the underlying emotion. The rule-based approach is only as good, as is its parsing algorithm, and the quality of the lexical resource used for the lookup. Emotion lexicons list emotion-bearing words and classify them to single or multiple emotional categories. Word embeddings, on the other hand, also take into account frequently co-occurring words that are semantically similar.

Emotion lexicons can be built from scratch. However, there exist good off-the-shelf solutions. One of the most popular being *WordNet-Affect* (Strapparava et al., 2004). These off-the-shelf solutions differ tremendously in terms of their number of entries. *WordNet-Affect* contains close to five thousand words, whereas *DepecheMood* (Liu and Zhang, 2012), another popular lexicon, contains more than thirty-five thousand words. Nonetheless, the quality of the lexicon is not solely dependent on its size. The vocabulary used for the lexicon also impacts its quality. Bandhakavi et al. (2017) argue that general-purpose lexicons such as *WordNet-Affect* perform not as good as domain-specific emotion lexicons. Therefore, a smaller domain-specific lexicon might yield in better results than a larger general-purpose lexicon. *LIWC-based lexicons* (Linguistic Inquiry and Word Count) are also

widely used, since these dictionaries list grammatical, psychological, and content word categories, and thus also emotion categories with thousands of entries (Chung and Pennebaker, 2012).

The idea behind word embeddings is similar to emotion lexicons. Each word is represented as a vector in the vector space. Thereby, frequently co-occurring words are considered semantically similar and therefore, close in the vector space (Seyeditabari et al., 2018). Among the most popular word embedding methods is *word2Vec* (Mikolov et al., 2013). Word embeddings are also often used to train machine learning models, like *LSTM*, which usually take word vectors as inputs.

Both rule-based approaches are straightforward. However, there are some drawbacks to them. The emotional meaning of keywords can be ambiguous and is context-sensitive. The sentences *She hates me*, and *I hate her*, could both be classified as *anger* based on the keyword *hate*. However, when looking at the sentence level information, the first utterance could also be perceived as *sad*. Ignoring the syntactic structure and semantics of the whole sentence, can therefore lead to misinterpretations. Furthermore, sentences without any emotional keywords cannot be classified. Even if they might contain an implicit expression of emotions, for example in the form of a metaphor (Kao et al., 2009). As a consequence, especially emotion lexicons often lack accuracy compared to more complex approaches.

2.2.2 Non-Neural Machine-Learning

Unlike rule-based approaches, non-neural learning-based approaches are trying to detect emotions using trained classifiers, such as the *Support Vector Machine (SVM)* (Teng et al., 2006), *Naive Bayes,* or *Decision Trees*.

We distinguish between *supervised* and *unsupervised* learning. Unsupervised approaches are an evolution of the rule-based approaches and are learning from test data that is not annotated with emotional labels. Most commonly, these approaches use movie dialogues (Banchs, 2017; Honghao et al., 2017) or children's fairy tales (Kim et al., 2010) to build emotional lexicons and train their models.

Supervised approaches, on the other hand, learn from labelled data such as Twitter messages. Common labels are annotations, hashtags or emojis. There exist a few good sources for emotionally labelled text, one of the most prominent being the *Swiss Center for Affective Sciences*[7] providing datasets like the *International Survey On Emotion Antecedents And Reactions (ISEAR)* and other useful tools for emotion detection. Other well-known datasets are *EmotiNet* (Balahur et al., 2011) and *SemEval-2007* (Strapparava and Mihalcea, 2007).

As stated by Seyeditabari et al. (2018), one of the major challenges for supervised approaches is the lack of quality training data. Oftentimes, these datasets are unbalanced in terms of emotion categories. Banchs (2017) analyse the large movie dialogue dataset *MovieDiC* and conclude that emotions such as *love*, or *joy* occur much more frequently than *fear* or *surprise*. The classifiers trained on such datasets will therefore underperform for these emotional categories.

2.2.3 Deep Learning Approaches

Most recent advances that showed to be effective in the field of emotion detection, have been made using deep learning (Xu et al., 2017).

Oftentimes, deep learning approaches are covering both, the emotion detection and the response generation, for example when using an *Encoder-Decoder architecture* (Serban et al., 2015). This architecture consists of two stages — the encoding and decoding stage. In the encoding stage, the raw text input is turned into a feature representation, usually in the form of a vector. The vector is then used as an input for the decoding stage to generate a response by applying the same strategies as in the encoding stage, but in the opposite direction.

A well-known approach applying the encoding-decoding architecture is *Long Short-Term Memory (LSTM)* (Jithesh et al., 2017). LSTM is a *Recurrent Neural Network (RNN)* that allows to capture long-term dependencies and store sequential information over a longer time. It can retain and forget the previous state and memorise extracted information from the input data depending on its importance (Xu et al., 2017; Sun et al., 2019).

The commonly used *Sequence to Sequence (Seq2Seq)* model also uses LSTMs and the Encoding-Decoding architecture (Sutskever et al., 2014). There is one LSTM for the encoding stage, transforming the raw text input into a fixed-length vector representation, whereas another LSTM is used for the decoding to a variable-length text out-

[7] https://www.unige.ch/cisa/

put (Cho et al., 2014; Xu et al., 2017; Chan and Lui, 2018).

To improve the model's efficiency, Chan and Lui (2018) investigate different approaches on embedding emotional information for Seq2Seq models. Different styles, positioning, and embeddings of emotional information are tested. The authors conclude that the positioning in general matters and impacts the emotion detection.

2.3 Response Generation

One of the most difficult tasks for empathic chatbots is generating an empathic response. Firstly, because it faces similar challenges as the emotion detection stage. Secondly, because it not only has to ensure that the response is appropriate in terms of content level information, but also in terms of emotion level information. This balancing act is tremendously difficult, as one usually has to sacrifice accuracy for one of the information levels, when trying to optimise the other (Xu et al., 2017; Zhou et al., 2017).

In terms of empathic response generation, we distinguish between two strategies — *retrieval-based approaches* and *dynamic generation*.

2.3.1 Retrieval-Based approaches

These approaches look up common responses to the user's utterance in conversation datasets. However, this method is very limited in its applicability. Similar inputs yield in the same responses, making the conversation repetitive and less natural. Furthermore, huge datasets of emotional conversations are required for such systems in order to achieve acceptable results. As such datasets are scare, these types of approaches tend to yield in responses such as *I don't know* in cases where no candidate response can be found.

A more advanced version of retrieval-based systems uses word embeddings on the input text to find the closest candidate responses, thus yielding in slightly more diverse responses (Bartl and Spanakis, 2017). However, it still requires lots of emotionally charged sample conversations.

Empathic response generation in general requires similar datasets as emotion detection, but with a bigger focus on conversation and dialogue turns. Movie dialogues are a good source for emotionally charged conversations. However, they oftentimes do not resemble daily conversation and seem more artificial and theatrical (Chan and Lui, 2018). Furthermore, emotions in movie dialogues are after all still acted and not naturally occurring, which might also have an impact on the quality of the training data.

Other common datasets include chat conversations, for example from Twitter service requests (Xu et al., 2017) that might yield in more natural conversations.

2.3.2 Dynamic Generation

These approaches are strongly tied with the deep-learning approaches used for emotion detection from Section 2.2.3 and usually based on the encoder-decoder architecture, such as the Sequence-to-Sequence model.

The input sentence is encoded on a word-by-word basis by embedding each word separately, whilst taking into account already encoded words using hidden states. Thereby, the last word embedding will produce a vector representation of the whole input sequence, encapsulating all relevant sentence level information. Semantically similar sentences are therefore close to each other in a vector space (Sutskever et al., 2014).

The decoder will then use the sentence vector or rather *sentence embedding* to produce an output sentence using inverted encoding mechanisms on a word-by-word basis. This allows encoding-decoding architectures to generate variable length responses. Thereby, it will consider already decoded words to ensure that the generated response is also grammatically correct. To find appropriate responses to a given word from the encoding stage, vocabularies or word-embeddings are being used.

However, the longer the input sequence, the more challenging to capture the full meaning in a single sentence embedding. For an input sentence of 30 words, the decoder would have to consider, what was encoded 30 steps ago, just to decode the first word. This long-range dependency problem oftentimes results in poor responses, like *I don't know* (Chan and Lui, 2018).

Attention mechanisms are commonly used to tackle the issue of long-range dependencies (Sutskever et al., 2014). Using attention, the decoder has direct access to the hidden state of each encoded word and can weight each word correspondingly. This allows the decoder to attend and weight on relevant parts of an input sentence, when generating the output. This mechanism is also applied in Neural Machine Translation (Bahdanau et al., 2014).

However, these Recurrent Neural Networks (RNN) still suffer from the vanishing gradient problem, that causes issues with long-range dependencies (Hochreiter, 1998). LSTMs also apply attention and in addition allow to retain and forget information, therefore handling the long-range dependency problem better than other approaches.

All these mechanisms are essentially required to ensure an appropriate response in terms of content level information. When we also want to consider emotion level information, we add additional complexity to the model. Emotions either have to be additionally encoded during the encoding stage or fed directly to the decoding stage to generate emotional responses (Zhou et al., 2017).

The Emotional Chatting Machine as proposed by Zhou et al. (2017) is a recent and noteworthy approach for assessing the emotional state of conversations and to generate appropriate emotional responses. Therefore, it belongs to the dynamic generation approaches. The ECM deep learning algorithm is trained with 22.300 Chinese blog posts that are manually annotated with Ekman's six basic emotions.

In terms of architecture, the ECM is based on the Seq2Seq model with an encoding and decoding phase. In addition, Zhou et al. (2017) introduce an internal and external memory to the model to capture changes in the emotion state throughout the sentence and to map explicit emotion expressions to emotion categories. Figure 2 provides a good overview of the ECM architecture. As input, the ECM requires the user's text message and one of the *Ekman's six emotion* categories to condition the response. Based on the input message, the ECM will generate an appropriate response and condition it using the input emotion category. Zhou et al. (2017) benchmark the ECM with other approaches, such as the traditional Seq2Seq model or lexicon-based approaches and show that ECM performs best across all emotion categories. However, it lacks behind slightly on the content level that the authors put down to an imbalance in the training set.

One drawback of the ECM is that the input emotion has to be set manually. This hardwiring of the output emotion can be useful, if the chatbot should always respond in the same emotional state, or express certain personality traits such as being angry all the time. However, if we want the chatbot to dynamically react to the user's emotions and make the interaction natural, then we have to change the emotion category based on the emotional state of the user's message automatically.

2.3.3 Empathic Responses

As discussed in Section 1, empathy requires understanding of the user's emotion and replying to them on an appropriate emotional level. Using emotion detection, we can achieve good results in understanding the user's emotions. The difficult part is actually selecting the *appropriate* emotion to condition the response with.

One approach could be to simply mirror the user's emotion. However, in human-conversations, empathy finds expression, when one tries to feel with the opponent and not necessarily similar to the opponent. One's own emotion must not be confused with the opponent's emotion. When resonating to the opponent's emotion, one is still aware that it might be different from the personal emotion (Singer and Klimecki, 2014). If someone is sad, you might understand this sadness and try to cheer them up, instead of responding in a sad way as well. However, it does not mean that you are necessarily feeling sad as well. Thus, simply mirroring the user's emotion does not necessarily yield in empathic responses.

This is also an important aspect with regards to a chatbot's personality, since in these cases, one should think about the chatbot's own emotion, as well as how it might resonate on someone else's emotions.

Another crucial aspect is taking into account the user's emotional evolution throughout the whole conversation. When just considering the latest user utterance for emotion detection, misinterpretations or frequent switches in the emotions expressed by the chatbot's response might occur. For example, the user could genuinely be in a bad mood, but laugh at a joke one just made. If we would consider just the latest user utterance to detect the user's emotion and condition the response accordingly, the chatbot's expressed emotion would switch from negative to positive within a single sentence. Modelling the user's emotion over a longer period might also be important when applying empathic chatbots in treating mental illnesses, or when building personality profiles to monitor the emotional state of the user.

We observe that only little research actually focusses on the generation of empathic responses in Computer Science, compared to the efforts

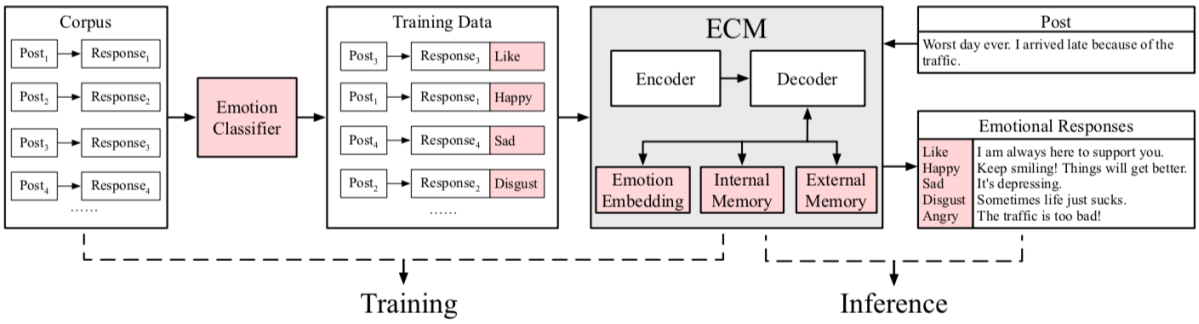

Figure 2: Overview of the ECM architecture based on the encoder-decoder framework with the addition of an internal and external memory to further improve the emotional response (Zhou et al., 2017).

done for emotion detection. There exist some approaches, such as the Encoder-Decoder architecture, that cover emotion detection as well as emotional response generation. Nonetheless, an emotional response is not necessarily an empathic response as elaborated before.

How humans are generating empathic responses is still an ongoing field of research in neuroscience (Shamay-Tsoory and Lamm, 2018). Similar to emotions, there is no universally accepted model for empathic responses, except that empathy is heavily context-dependent (Singer and Klimecki, 2014).

2.4 Response Expression

In a normal conversation, non-verbal cues such as facial expressions or gestures can help indicate a person's emotional state. However, with chatbots, we are missing such information and have to focus solely on the user's text, to detect the emotional state. Similar constraints also apply to the response expression by the chatbot. It is difficult to transport the intended emotion from the generated response back to the user in a text format. Some approaches try to simulate non-verbal cues by displaying the chatbot as a 3D simulation of a person (Tatai et al., 2003). We note that in general, chatbots do not express responses in any other way than text. Because such non-verbal cues are missing in traditional electronic messaging systems, people are using emojis to supply such cues. The usage of emojis has increased heavily over the previous years. In 2017, Facebook revealed that on an average day, over 5 Billion emojis are being sent over Messenger only[8]. Hu et al. (2017) state that the main reason for the usage of emojis in messages is to express emotions or strengthen expressions.

Emojis could therefore also be considered when detecting the user's emotion. However, two challenges arise when using emojis as possible emotional labels. First, the emoji label could be contradictory to the perceived emotional state from the text, for example implying a sarcastic utterance. Figure 3 shows, how emojis can lead to such contradicting interpretations. Second, emojis are prone to cultural differences as stated by Ljubešić and Fišer (2016). Chatbots with a global scope should therefore take into account, that emojis might be used and perceived differently depending on the country.

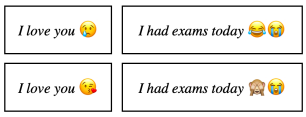

Figure 3: Two examples of challenging cases where the emojis are contradictory to the perceived emotional state of the text message, or the multiple contradicting emojis are used.

DeepMoji (Felbo et al., 2017) is an impressive tool translating text into a set of emojis expressing a similar emotional state returning the five most likely emojis together with their probabilities. It is trained on 1.2 billion tweets containing emojis and uses LSTM to predict the most appropriate emojis. Generated emojis could be mapped to different emotion categories and used to express emotions in the response to the user. We leave the validation of this method for future work.

We state that the usage of emojis in conversational agents might be a clue to make them more human-like and to help expressing non-verbal cues

[8] https://www.adweek.com/digital/facebook-world-emoji-day-stats-the-emoji-movie-stickers/

that otherwise might go missing. Future work should therefore focus on validating this hypothesis.

3 Discussion and Conclusion

We note that current state-of-the-art approaches face the following major challenges:

1. Shortage of quality training data — Machine-Learning algorithms for emotion detection and empathic response generation require an extensive amount of annotated training data. Existing datasets are scarce and are oftentimes unbalanced for different emotions. Hence, chatbots that were trained using such datasets will lead to poor performances on these emotions. Using annotated data from social media has proven to yield in good results, but also suffers from unbalanced emotion distribution. To generate human-like responses, natural conversations should be used for training as opposed to artificial and theatrical movie dialogues.

2. Emotion level and content level — To generate responses that are grammatically correct and that reflect the appropriate content level is very complex. Using domain specific training data can help improve the accuracy. If the answer should also reflect the emotional level and detect possibly implicit or multi-layered emotions, then the complexity increases even further. Improving one of the levels — emotion or content — without sacrificing accuracy for the other is very challenging. For response generation, we note that existing approaches are mainly focusing on content level information and consider emotions only as additional information during encoding.

3. Considering the full end-to-end experience — In order to achieve good results, one has to consider the impacts of all four stages — emotion expression by the user, emotion detection by the chatbot, response generation by the chatbot, and appropriate response expression back to the user. Only by considering the full end-to-end experience can chatbots be improved to be more human-like and empathic.

Future work should investigate the integration of emojis into the full end-to-end experience — from emotion detection to response expression.

4. Modelling emotions throughout conversations — We state that when selecting emotions to condition the response, one should not only consider the detected emotion from the latest user message. Taking into account the evolution of the user's emotion throughout the whole conversation and possibly even over several previous conversations, prevents frequent changes of the chatbot's expressed emotions and helps model the user's long-term emotional state.

Furthermore, more efforts should be devoted to understanding empathy and how chatbots can generate empathic responses instead of just emotional responses.

In this survey, we have presented the state-of-the-art of empathy and especially empathic response generation in chatbots and pointed out several noteworthy approaches. We pointed out the four stages of the interaction with the chatbot and underlined the importance to take all the stages into account when creating empathic chatbots.

We note that there exist many different approaches to tackle the problem of emotion detection, but only few for empathic response generation. Overall, deep learning algorithms, such as the Emotional Chatting Machine (ECM) tend to yield in the best results. Even though, there is still potential for improvement as the ECM only generates emotional responses, but not empathic ones.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* https://arxiv.org/abs/1409.0473.

Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. 2011. EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories. In *Natural Language Processing and Information Systems*, Springer Berlin Heidelberg, pages 27–39. https://doi.org/10.1007/978-3-642-22327-3$_4$.

Rafael E. Banchs. 2017. On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. https://doi.org/10.1109/apsipa.2017.8282245.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. 2017. Lexicon generation for emotion detection from text. *IEEE Intelligent Systems* 32(1):102–108. https://doi.org/10.1109/mis.2017.22.

A. Bartl and G. Spanakis. 2017. A retrieval-based dialogue system utilizing utterance and context embeddings. In *2017 16th IEEE International Conference on Machine Learning*

and Applications (ICMLA). pages 1120–1125. https://doi.org/10.1109/ICMLA.2017.00011.

Aaron Ben-Zeev. 2000. *The Subtlety of Emotions*. The MIT Press. https://doi.org/10.7551/mitpress/6548.001.0001.

Yin Hei Chan and Andrew Kwok Fai Lui. 2018. Encoding emotional information for sequence-to-sequence response generation. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE. https://doi.org/10.1109/icaibd.2018.8396177.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1179.

Cindy K. Chung and James W. Pennebaker. 2012. Linguistic inquiry and word count (LIWC). In *Applied Natural Language Processing*, IGI Global, pages 206–229. https://doi.org/10.4018/978-1-60960-741-8.ch012.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6(3-4):169–200. https://doi.org/10.1080/02699939208411068.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d17-1169.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health* 4(2):e19. https://doi.org/10.2196/mental.7785.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6(02):107–116. https://doi.org/10.1142/S0218488598000094.

WEI Honghao, Yiwei Zhao, and Junjie Ke. 2017. Building chatbot with emotions http://web.stanford.edu/class/cs224s/reports.

Philip N. Howard, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2018. Social media, news and political information during the US election: Was polarizing content concentrated in swing states? *CoRR* abs/1802.03573. http://arxiv.org/abs/1802.03573.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. 2017. Spice up your chat: The intentions and sentiment effects of using emoji. *arXiv preprint arXiv:1703.02860* https://arxiv.org/abs/1703.02860.

V Jithesh, M Justin Sagayaraj, and K G Srinivasa. 2017. LSTM recurrent neural networks for high resolution range profile based radar target classification. In *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*. IEEE. https://doi.org/10.1109/ciact.2017.7977298.

Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. 2009. Towards text-based emotion detection a survey and possible improvements. In *2009 International Conference on Information Management and Engineering*. IEEE, pages 70–74. https://doi.org/10.1109/icime.2009.113.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A. Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Stroudsburg, PA, USA, CAAGET '10, pages 62–70. http://dl.acm.org/citation.cfm?id=1860631.1860639.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, Springer US, pages 415–463. https://doi.org/10.1007/978-1-4614-3223-4$_1$3.

Bingjie Liu and S. Shyam Sundar. 2018. Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21(10):625–636. https://doi.org/10.1089/cyber.2018.0110.

Nikola Ljubešić and Darja Fišer. 2016. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop*. Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-2610.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* https://arxiv.org/abs/1301.3781.

Keith Oatley, Dacher Keltner, and Jennifer M Jenkins. 2006. *Understanding emotions*. Blackwell publishing.

W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.

Robert Plutchik. 1991. *The emotions*. University Press of America.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR* abs/1507.04808. http://arxiv.org/abs/1507.04808.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674* https://arxiv.org/abs/1806.00674.

Simone Shamay-Tsoory and Claus Lamm. 2018. The neuroscience of empathy – from past to present and future. *Neuropsychologia* 116:1 – 4. Special Issue: The Neuroscience of Empathy. https://doi.org/10.1016/j.neuropsychologia.2018.04.034.

Tania Singer and Olga M. Klimecki. 2014. Empathy and compassion. *Current Biology* 24(18):R875 – R878. https://doi.org/10.1016/j.cub.2014.06.054.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval 07*. Association for Computational Linguistics. https://doi.org/10.3115/1621474.1621487.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*. Citeseer, volume 4, pages 1083–1086.

Xiao Sun, Chen Zhang, and Lian Li. 2019. Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor. *Information Fusion* 46:11–22. https://doi.org/10.1016/j.inffus.2018.04.001.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR* abs/1409.3215. http://arxiv.org/abs/1409.3215.

Gábor Tatai, Annamária Csordás, Árpád Kiss, Attila Szaló, and László Laufer. 2003. Happy chatbot, happy user. In *Intelligent Virtual Agents*, Springer Berlin Heidelberg, pages 5–12. https://doi.org/10.1007/978-3-540-39396-2_2.

Zhi Teng, Fuji Ren, and Shingo Kuroiwa. 2006. Retracted: Recognition of emotion with SVMs. In *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pages 701–710. https://doi.org/10.1007/978-3-540-37275-2_87.

Samuel C Woolley and Douglas R Guilbeault. 2017. Computational propaganda in the united states of america: Manufacturing consensus online. *Computational Propaganda Research Project* page 22. http://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2017/06/Comprop-USA.pdf.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI 17*. ACM Press. https://doi.org/10.1145/3025453.3025496.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074* https://arxiv.org/abs/1704.01074.