SarcEmp: Fine-tuning DialoGPT for Sarcasm and Empathy

Mohammed Rizwan
Independent Researcher
Kota, Rajasthan, India
mohrizwan89.rq@gmail.com

Abstract

Conversational models often face challenges such as a lack of emotional temperament and a limited sense of humor when interacting with users. To address these issues, we have selected relevant data and fine-tuned the model to (i) humanize the chatbot based on the user's emotional response and the context of the conversation using a dataset based on empathy and (ii) enhanced conversations while incorporating humor/sarcasm for better user engagement. We aspire to achieve more personalized and enhanced user-computer interactions with the help of varied datasets involving sarcasm together with empathy on top of already available state-of-the-art conversational systems.

1 Introduction

Recent advancements in large-scale pre-trained models, such as models using transformer-based architectures, have produced impressive results, as seen with DialoGPT (Yizhe Zhang, 2020). However, it is only recently that these models have had access to enough data to respond in a neutral tone and provide information based solely on the user's input. Understanding the emotional response and situation of the user is not an easy task, especially when it comes to providing an appropriate response. Early sarcasm detection methods heavily depended on static textual patterns like lexical indicators, syntactic rules, and specific emoji occurrences (Dmitry Davidov, 2010), (Maynard and Greenwood, 2014), (Bjarke Felbo, 2020). Unfortunately, these methods often under-performed and lacked generalization due to their inability to leverage contextual information effectively. Additionally, they faced issues with poor performance. Another problem is the lack of an explicit long-term memory of the conversation because these systems are trained to generate a response based only on the recent dialogue history (Oriol Vinyals, 2015) [A neural conversational model]. Recently, chatbots have faced challenges in providing inaccurate, nonsensical, or insensitive responses, largely stemming from a lack of contextual understanding and emotional awareness during conversations.

The key aim of this work is to enhance the neutral persona-based models by incorporating sarcasm and an empathetic touch. To achieve this, we fine-tuned the DialoGPT model using two datasets. These datasets include 1.3 million sarcastic comments from Reddit and 25,000 personal dialogues in which a speaker expressed a specific emotion, and a listener responded.

2 Related Work

Recent developments in engaging dialogue agents with a 'profile' (Saizheng Zhang, 2018) have helped the vision of contextually aware chatbots immensely. This allows models to respond by sticking with a persona, and hence, replies are more stable and coherent. Research on the emotional spectrum, including models that incorporate a sense of humor or sarcasm, is still being refined due to the challenging task of understanding the nuanced nature of sarcasm. According to the Khatri et al. (2018), sarcasm can be difficult to detect and harder to eradicate because abuse is sometimes hidden behind it. It will take much progress in the field to detect and generate sarcasm accurately. There have been models developed with the ability to detect sarcasm from user input (Devin Pelser, 2019). However, generating responses in the same fashion is not yet fully addressed. Even models with the ability to produce empathetic responses (Hannah Rashkin, 2019) do not fully capture the wide range of emotions experienced by a typical human being and respond accordingly. Another problem is the lack of an ex-

Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024), pages 66–69

plicit long-term memory of the conversation. Typically, these systems are trained to generate a response based only on the recent dialogue history (Oriol Vinyals, 2015).

3 Methodology

3.1 DialoGPT

DialoGPT is well-suited for fine-tuning with multiple datasets due to its versatile architecture and pre-training on a diverse range of conversational data. It is based on GPT-2 (Alec Radford, 2018) architecture, making user-specific prompts more realistic. DialoGPT employs maximum mutual information (MMI) scoring function (Saizheng Zhang, 2018), integrating a pre-trained backward model. This model predicts source sentences from responses and filters out bland or uninformative text, ensuring it generates contextually relevant and meaningful responses. MMI enhances the model's ability to avoid generic replies, making its conversations more engaging and purposeful. The model also exhibits the capability to address commonsense questions to some extent, due to the rich amount of information learned from Reddit data. It also shows consistency with respect to the context in multi-turn generation, outperforming RNN counterparts and tending to be more consistent with the context. Additionally, the release of the source code and pre-trained models facilitates future research and development, providing a foundation for novel applications and methodologies. Furthermore, the model's performance in the (Yoshino et al., 2019) DSTC-7 Dialogue Generation Challenge demonstrates its potential for generating conversation responses grounded in external knowledge, making it suitable for applications requiring information-rich interactions. Its ability to surpass human responses in automatic metrics also indicates its potential for enhancing human-computer interactions in various domains. Fine-tuning DialoGPT can lead to the development of more intelligent open-domain dialogue systems tailored to specific contexts or domains.

3.2 Datasets

To fine-tune the Dialo-GPT model, we have used two datasets to achieve our target. We explain the datasets in the following subsections.

3.2.1 SARC

The Self-Annotated Reddit Corpus (SARC), is a significant resource for sarcasm research and the development of systems for sarcasm detection (Mikhail Khodak, 2018). It addresses the challenge of detecting sarcasm in natural language processing, emphasizing the difficulty in discerning sarcasm due to its infrequent occurrence and complexity. The SARC dataset comprises 1.3 million self-annotated sarcastic statements, surpassing previous datasets in size by an order of magnitude. This large corpus provides opportunities for balanced and unbalanced label learning, enabling the evaluation and training of sarcasm detection systems.

We'll fine-tune DialoGPT for generating sarcastic text using the SARC dataset, comprising self-annotated sarcastic statements containing '/s'(sarcasm tag). This dataset includes conversation threads, responses, and sarcasm labels, serving as a benchmark for classifying statements. It comprises three essential components: the "label" indicating sarcasm or non-sarcasm, the "context" representing the parent comment preceding the response, and the "response" itself, serving as the answer to the preceding comment. By offering balanced learning tasks and methods for reducing false negatives, the SARC dataset aims to enhance machine learning methods and improve sarcastic text generation. It is freely available, fostering future research and the development of more effective sarcasm-based text generation and detection.

3.2.2 Empathetic Dialogues

The dataset, Empathetic Dialogues (Hannah Rashkin, 2019), is designed to serve as a new benchmark and training resource for evaluating the ability of dialogue models to generate empathetic responses. It is specifically tailored to address the challenge of empathetic responding, which involves recognizing and acknowledging the emotional cues and experiences expressed by a conversation partner in a dialogue. The dataset is best suited for training and evaluating dialogue systems, including chatbots and conversational agents, in their capacity to appropriately respond to personal experiences and emotions expressed in a conversation. This is a perfect dataset to be worked upon because it is a one-on-one conversation between a "Speaker" and a "Listener". The Speaker initiates a conversation by describing a situation, and the Listener becomes aware of it

Dataset	Perplexity		F1		Loss		Token Accuracy	
	SarcEmp	DialoGPT	SarcEmp	DialoGPT	SarcEmp	DialoGPT	SarcEmp	DialoGPT
Empathetic Dialogues	101.1	100.7	0.12	0.76	4.61	4.61	0.24	0.26
ConvAI2	141.6	144.6	0.08	0.78	4.95	4.97	0.18	0.18
Daily Dialogs	61.46	61.54	0.7588	0.05856	4.118	4.12	0.3328	0.2953

Table 1: Automatic metrics calculated on 1000 random examples from the mentioned datasets.

through the Speaker's words. Subsequently, the Speaker and Listener engage in six more additional turns (total 7 conversations). In each turn, a new emotion is given as a context, prompting the Listener to respond accordingly. These emotions consist of sentimental, afraid, proud, faithful, terrified, joyful, and angry.

3.3 Fine-tuning Process

Upon acquiring our datasets, the initial step involves processing the data to align with the model's comprehension. In the SARC dataset, comments labeled as "Sarcastic" are initially sorted based on their labels. Subsequently, this sorted data is formatted to feed into two distinct fields: 'context' and 'response,' ensuring compatibility with the model's understanding. Here, the parent comment takes the place of 'context,' while the corresponding response is assigned to the 'response' field.

Within the Empathetic Dialogues dataset, the information is structured around 'prompts' and 'utterances.' The 'prompt' signifies a sentence that is awaiting refinement based on the context, while 'utterance' represents the corresponding response aligned with that context. In this dataset, the 'prompt' is mapped to the 'context' field, and the associated 'utterance' is placed in the 'response' field for compatibility with the model's understanding.

The next step involves the concatenation and randomization of all 'response' and 'context' pairs using the Pandas Library. Subsequently, the entire dataset is divided into training (60%) and testing (40%) segments. To ensure model comprehension, each row's data is combined into a single string. A special 'end of string' token is inserted between individual strings, facilitating the model in recognizing the conclusion of each response within the string. This process streamlines the dataset for effective training and testing, enhancing the model's ability to understand and generate responses. Following the concatenation and randomization process, the data undergoes tokenization and is subsequently trained using checkpoints and a set number of epochs. The objective here is to fine-tune the model and evaluate its perplexity and other automatic metrics. The perplexity of a model entails evaluating the model's predictive accuracy in determining the next token within the sequence. The incorporation of checkpoints aids in the continuous monitoring and preservation of the model's progress during the training process, ensuring optimal performance.

4 Results

In this project, we choose perplexity as an automatic metric to evaluate the model's performance among others such as f1 score, loss, and token accuracy. Perplexity loss measures how well a model can predict the next word in a sequence of text. Lower values indicate a better understanding of the language and context. Perplexity is also reported to have a robust correlation with human perceptions of coherent and contextually specific natural conversations (Adiwardana and et al., 2020). We report the results in the table 1.

The fine-tuning of DialoGPT on the SARC and Empathetic Dialogues datasets has yielded noteworthy results. The model achieves a low training perplexity of 2.996, showcasing improved confidence in predicting the next token. We also perform a variety of experiments by training and evaluating different models. Table 1 depicts the performance of our model SarcEmp across three major datasets, including empathetic dialogues (Hannah Rashkin, 2019), ConvAI2 (Dinan et al., 2019), and Daily Dialogues (Li et al., 2017). We can observe that the fine-tuned model performs similarly to the baseline model and even better in some cases. Perplexity is reportedly reduced in two of the tested datasets, while loss on the datasets is consistently low. However, the difference is not significant when it comes to automatic metrics, but it will be interesting to observe the human evaluation results in a more realistic setting.

5 Conclusion

In this work, we have introduced a new fine-tuned model that incorporated sarcasm and empathy on top of a state-of-the-model. The resulting model is performing pretty consistently in terms of automatic evaluation, although it would be interesting to see it perform in human evaluation tasks. We believe human evaluation would provide us with useful insights into the domain of more engaging and empathetic human-computer interactions and potential directions for improvements.

References

Michel Galley et.al. Yizhe Zhang, Siqi Sun. Dialogpt : Large-scale generative pre-training for conversational response generation. 2020.

et al. Dmitry Davidov, Oren Tsur. Semi-supervised recognition of sarcastic sentences in twitter and amazon. 2010.

Maynard and Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. 2014.

Anders Søgaard et al. Bjarke Felbo, Alan Mislove. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. 2020.

Quoc Le. et al. Oriol Vinyals. A neural conversational model. 2015.

Jack Urbanek et al. Saizheng Zhang, Emily Dinan. Personalizing dialogue agents: I have a dog, do you have pets too? 2018.

Chandra Khatri, Behnam Hedayatnia, and Rahul Goel et al. Detecting offensive content in open-domain conversations using two stage semi-supervision, 2018.

Hugh Murrell Devin Pelser. Deep and dense sarcasm detection. 2019.

Margaret Li et al. Hannah Rashkin, Eric Michael Smith. Towards empathetic open-domain conversation models: a new benchmark and dataset. 2019.

Tim Salimans Ilya Sutskever Alec Radford, Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda Alamari, Tim K. Marks, Devi Parikh, and Dhruv Batra. Dialog system technology challenge 7, 2019.

Kiran Vodrahalli Mikhail Khodak, Nikunj Saunshi. A large self-annotated corpus for sarcasm. 2018.

Daniel Adiwardana and Minh-Thang Luong et al. Towards a human-like open-domain chatbot, 2020.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The second conversational intelligence challenge (convai2), 2019.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset, 2017.