

Emotion-and-knowledge grounded response generation in an open-domain dialogue setting

Deeksha Varshney^a, Asif Ekbal^a, Erik Cambria^{b,*}

^a Department of Computer Science and Engineering, IIT Patna, India

^b School of Computer Science and Engineering, NTU, Singapore

ARTICLE INFO

Keywords:

Response generation
Emotion
Knowledge
Transformers

ABSTRACT

The neural-based interactive dialogue system focuses on engaging and retaining humans in long-lasting conversations. This has been explored for a variety of goal-oriented dialogue domains, such as education, health care, entertainment, sports, and politics. To develop an understanding and awareness of social and cultural norms, and to address specific social skills, we need to invent strategies for building interactive systems that take into account the user's emotions and relevant-facts in a multi-turn conversation. In this paper, we propose a new neural generative model that combines step-wise co-attention with a self-attention-based transformer network along with an emotion classifier to jointly control emotion and knowledge transfer during response generation. Quantitative, qualitative, and human evaluation results on the benchmark Topical Chat and the CMU_DoG dataset show that the proposed models can generate natural and coherent sentences, capturing essential facts with considerable improvement over emotional content.

1. Introduction

Chatbots are on the rise due to their ability to mimic conversations on several topics which often require social, emotional, and cognitive skills to encourage interactive conversations with the target audience. In recent years, there has been an increasing focus on creating neural-based conversational systems that operate in an open-domain setting. Generative neural networks that use Seq2Seq mechanisms, such as those described in previous studies [1–4], often struggle to accurately mimic human conversation. They tend to produce uninteresting, monotonous, and formulaic responses, such as “I’m not sure”, “That’s interesting”, and “That’s a good question”.

Therefore, for many natural language generation (NLG) tasks, semantic understanding is crucial, particularly when it is supported by real-world knowledge or commonsense. To create an effective open-domain dialogue generation system, it is necessary to incorporate the pertinent themes, objects, and connections referenced in the user's input while generating an appropriate response. Unlike structured databases, conversational datasets require the ability to comprehend relevant knowledge from a range of sources, such as Wikipedia documents and news articles. For instance, as illustrated in Table 1, each statement in the conversation is linked to several sentences that provide information related to the conversation's subject from various sources, including Wikipedia.

Agent 2's response, “Apparently, 80 percent of the earth's natural forests have already been destroyed. It is depressing to me”, uses the unstructured data available to it. The objective here is not task completion as in traditional dialogue systems, but to build a high-quality information-seeking conversational system. Previous research efforts on building knowledge-driven dialogue systems have primarily focused on augmenting the grounded knowledge associated with each utterance to the input of the traditional Seq2Seq model [5,6]. In addition to grounding conversations with real-world knowledge, conversational datasets now incorporate a variety of other types of information, including human emotions. The user's emotional state is a critical indicator for the system to produce more satisfying, intelligent, and social responses.

It is essential to generate responses that not only align with the topic of the user's input but also address and validate any emotions expressed in the user's query. For instance, in Table 1, the response “That is sad for sure. Sad that people are the cause of it too” is more rewarding because it attends to users' inherent feelings and emotions in a more affecting way. An emotionally intelligent conversational agent can improve the user experience by providing them with an interesting conversation that can effectively acknowledge their emotional needs. [7] introduces EmoKbGAN, a novel conversational model that harnesses a knowledge base and emotion labels to craft engaging dialogues. Comprising a transformer-based language model and twin discriminators, it oversees the generation procedure.

* Corresponding author.

E-mail address: cambria@ntu.edu.sg (E. Cambria).

Table 1

A section of insightful conversation and its related knowledge sentences extracted from the Topical Chat dataset. The sentence selected from the relevant knowledge base is emphasized in bold.

	Utterance	Knowledge
Agent 1	Let's talk about Earth. It's the third planet from the sun.	During this time, Earth rotates about its axis about 366.26 times. Earth's axis of rotation is tilted with respect to its orbital plane, producing seasons on Earth. ...
Agent 2	Apparently, 80 percent of the earth's natural forests have already been destroyed. It's depressing to me.	more than 80 percent of the Earth's natural forests have already been destroyed But I haven't seen it, ...
Agent 1	That is sad for sure. Sad that people are the cause of it too.	Sure, the idea that NASA would send oil drillers to bore into an asteroid propelling toward Earth may seem insane. ...
Agent 2	When the earth was first formed, there was only 5.5 h long in a day.	when Earth was first formed a day was only 5.5 h long

Instead of the traditional maximum likelihood estimation, they advocate for multi-attribute discriminator training to guide the learning. Although such advancements have been pivotal, existing models like EmoKbGAN often have limitations in handling vast unstructured knowledge bases, particularly when trying to assimilate dynamic conversational contexts. In contrast to these antecedent endeavors, our approach pivots on the co-attention architecture to adeptly manage unstructured databases like Wikipedia within dialogues.

This study introduces a new approach for constructing a neural conversation model based on knowledge and emotional labels, which aims to sustain extended conversations with users. The encoder utilizes a sequential positional self-attention-based transformer framework with a step-wise co-attention mechanism to encode the multi-turn dialogues. It consists of two positional self-attention blocks that replace Long Short Term Memory (LSTM) for modelling data dependencies, and a context-knowledge co-attention block to simultaneously attend to both context and related knowledge, thereby effectively integrating multiple knowledge sources into the multi-turn conversations. The decoder is motivated by the prior research of [8] that generates the target response in two phases. The first phase involves a decoder that generates responses by taking into account the context, whereas in the second phase, the decoder considers the knowledge for generation. We further enhance the response generation step by adding a binary classifier which is used as an additional module to guide the training process for emotional dialogue generation.

We evaluate the novel model on both the Document Grounded Conversation dataset [9] and the Topical Chat dataset [10]. We show through our experiments that the generated responses from our approach are capable of inducing emotions and preserving much contextual information. We find that controlling both the knowledge and emotional attributes of a conversation contributes significantly towards the development of highly interactive social chatbots and intelligent dialogue systems. Our findings demonstrate that the suggested approach significantly surpasses the benchmark methods in terms of performance, as evaluated through both automatic and manual means.

The main highlights and characteristics of this work are outlined below:

1. We introduce a co-attention-based dialogue system that seamlessly integrates emotions and knowledge for captivating, topic-aware responses. Our system employs a bi-directional attention network, bridging knowledge-to-utterance and vice versa, optimized for unstructured databases. Additionally, a binary classifier conditioned on emotion labels steers the emotion-centric response generation.
2. The training objective of our model involves minimizing the combined losses from both our co-attention module and the emotion classifier.
3. The emotion classification task is accomplished using a BERT-based model trained on the Topical Chat dataset. The utterances in the Document grounded dataset are then annotated in a semi-supervised manner using this model.
4. Our extensive experiments and evaluations illustrate that our suggested approach efficiently employs real-world knowledge and the user's emotions to produce more dynamic and captivating dialogues.

2. Related work

Conventional conversational Artificial Intelligence (AI) systems can be grouped into two broad categories, *viz.* task-oriented [11–14] and open-domain [3,15–17]. Task-oriented dialogue systems are useful when we want to accomplish a specific task by conversing with users using dialogues of shorter lengths. In contrast, open-domain dialogue systems aim to create in-depth and captivating conversations with people about a wide range of topics. The widespread availability of conversational data across various domains has contributed to the significant rise in the popularity of chatbots in recent times [18]. Chatbots are becoming a must-have asset for leading businesses as they are designated to copy human conversations and master social skills that humans have in them naturally like recalling from world knowledge, explainable reasoning, and constructing valid arguments to have a smooth flow of conversation between a user and a bot.

Recently, there has been a surge in the development of open-domain dialogue systems that are capable of engaging in natural conversations with humans on a wide range of subjects, including education, healthcare, politics, and more. These chatbots have gained popularity and have found success in various real-world applications, such as customer assistance, social assistance, and personal assistance. Early chatbots were built using rule-based dialogue systems, such as ELIZA [19] or template-based systems like NJFun [20], Partially Observable Markov Decision Process (POMDP) [21], and statistical methods like Maximum Entropy Markov Models (MEMMs) [22]. However, these methods are not scalable and time-consuming, making them unsuitable for building chatbots for different domains and applications. Due to the emergence of online communication and the increasing popularity of messaging apps and chatbots, a large amount of conversational data is now available for research purposes. This has led to the development of retrieval-based [23–26] and neural-based [3,16,27–29] methods for the data-driven conversations.

The capabilities of conversational agents can be enhanced by incorporating external knowledge into traditional sequence-based models. [30] created a large conversation dataset grounded on knowledge retrieved from Wikipedia. They demonstrated convincing experiments which proved the usage of knowledge in generating naturally fluent responses. [6] conditioned replies on conversation history and external information to broaden the widely used Seq2Seq technique.

Memory networks and various GAN (Generative adversarial network) based techniques were used to jointly manage dialogue state and knowledge bases [26,31]. However, they did not particularly address the challenge of selecting the most relevant knowledge for conversation generation. This gap highlights the need for models or mechanisms that can adeptly choose pertinent knowledge to support more informed and contextually appropriate dialogues. Since choosing the relevant knowledge is a prerequisite for the success of knowledge-based dialogue systems, [32] proposes a Knowledge Selection method for knowledge-grounded conversation generation. It first calculates the discrepancy between the previously chosen information and the candidate knowledge sentences presented at the current turn. The end-to-end model includes the differential information with or separated from the contextual information to assist in the selection of the final knowledge. The development of an unsupervised method that jointly trains the learning of knowledge selection module along with fine-tuning of the pre-trained model was proposed in [33].

These works exhibit reliance on ample data, which may pose challenges in low-resource settings where such data is scarce. In low resource settings, DRD [34] learns to model dialogue using large-scale unstructured text by first employing two encoders to encode context and knowledge respectively and then uses a disentangled decoder in order to independently process the language, context, and knowledge. Extending the previous work, KAT [35] chooses the right knowledge to be used in generation by first comparing the candidate knowledge presented at the current turn with the previously selected knowledge. The ZRKG model [36] suggests a double latent variable model that shows both the knowledge relating to a context and response as well as the manner in which the knowledge is conveyed under zero resource situations, where no context-knowledge-response triples are accessible during training.

Understanding various affecting information, such as emotion, sentiment, empathy, etc. is crucial to building robust and human-like conversational systems. In recent times, quite a few approaches have been developed to monitor the emotional aspect while generating the relevant text. The generation of emotionally controlled text can be achieved by manually specifying the target emotion label [37–42]. For instance, the work by [43] proposed Affect-LM, an extension to LSTM for generating emotionally-tinted conversational text across four affect categories. The model allows customization of emotional content in generated sentences, evaluated positively in perception studies for natural emotional sentence generation and grammatical correctness. However, it relies on affective keyword spotting for training, potentially limiting generalization beyond predefined emotional keywords. To generate text for a given sentiment or emojis as the target labels, researchers have proposed the use of variational autoencoders (VAEs) [40] and conditional variational autoencoders (CVAEs) [38]. These methods are evaluated by matching the generated emotion with a fixed target emotion. Existing neural conversational models primarily focus on lexico-syntactic aspects, overlooking the affective content crucial in human dialogues. To address this, [44] utilized a continuous, three-dimensional representation of emotions, moving away from the manual specification of target emotion labels.

[45] proposed models for generating text with desired styles by either concatenating the desired emotion with the source input at the time of learning or incorporating emotion in the decoder. More sophisticated models, such as ECM (Emotional chatting machine) [37], extended the Seq2Seq framework by modelling the emotional factor of a sentence using three mechanisms, viz. emotion label embedding, external memory, and internal memory. However, both approaches had the shortcoming of requiring external input or decision-making in determining the emotion to be expressed. [46] provides a complete generative empathetic chatbot built with the help of the Generative Pre-trained Transformer (GPT), which can identify user emotions and respond sympathetically using transfer learning.

Further advancements in empathetic dialogue models [47], which utilizes structured knowledge such as commonsense knowledge and emotional lexical knowledge to construct an emotional context graph. The decoder learns the emotional dependencies using a multitasking framework, utilizing the cross-attention mechanism of the graph-aware transformer and the knowledge-enriched context. Neural style transfer in images is a well-studied problem. Motivated by images, there is a growing interest in developing style transfer algorithms for texts. Binary classifiers are commonly utilized as discriminators, providing a way to transfer properties of a sentence in some domain according to the sentences in the target domain. Using style-based discriminators, [12,48] introduced the stylistic transfer of consumer behaviours, such as courteousness. [49] used a language model as a structured discriminator to allow an efficient and persistent word-level assessment during the training process.

[50] provides an overview of the literature on empathetic dialogue systems. The review identifies three key features that underpin such systems: emotion awareness, personality awareness, and knowledge accessibility. These features allow the system to understand and respond to users in a more empathetic way. In a recent study by [51], the capabilities of ChatGPT, an emerging general artificial intelligence model, were evaluated for text classification on three affective computing problems - big-five personality prediction, sentiment analysis, and suicide tendency detection. The study utilized three baselines, including a robust language model (RoBERTa-base), a legacy word model with pre-trained embeddings (Word2Vec), and a simple bag-of-words baseline (BoW). The results showed that while the RoBERTa trained for a specific downstream task generally had superior performance, ChatGPT still provided decent results that were relatively comparable to the Word2Vec and BoW baselines. Additionally, ChatGPT demonstrated robustness against noisy data, whereas Word2Vec models achieved worse results due to noise. The study concludes that while ChatGPT is a good generalist model capable of achieving good results across various problems without specialized training, it is not as good as a specialized model for a downstream task.

[7] presents a novel conversation model called EmoKbGAN, which utilizes a knowledge base and emotion labels to generate more engaging responses. The model consists of a transformer-based language model and two discriminators that guide the generation process. They propose multi-attribute discriminator training as a replacement for the maximum likelihood estimation objective to supervise the training process. Different from these prior works, we leverage upon the co-attention architecture to handle the unstructured knowledge bases, such as Wikipedia documents in conversations with respect to the following two aspects: (i) using a bi-directional attention network based on a shared similarity matrix in two directions, viz. knowledge-to-utterance as well as utterance-to-knowledge for linking and fusing information from both the knowledge and the utterance words; and (ii) exploiting hierarchical encoding [4] architecture to encode the multiple turns in the conversation.

In our current work, we develop methods for efficiently modelling other attributes that are associated with a conversation. In particular, we build an intelligent dialogue system that generates responses conditioned on emotion and knowledge attributes and hence yields highly proactive conversations between the agent and the user. We use a self-attention-based transformer with a sequential co-attention model to encode multi-turn utterances in a conversation. We add a binary classifier, which is used as an additional module to guide the training process for emotional dialogue generation. We perform our experiments on two publicly available datasets, the knowledge-grounded Topical Chat dataset [10] and Document Grounded Conversation [9], a large open-domain conversational dataset. Our method shows that we can produce more varied and human-like responses.

3. Methodology

In this section, we first define the problem and then describe the proposed method.

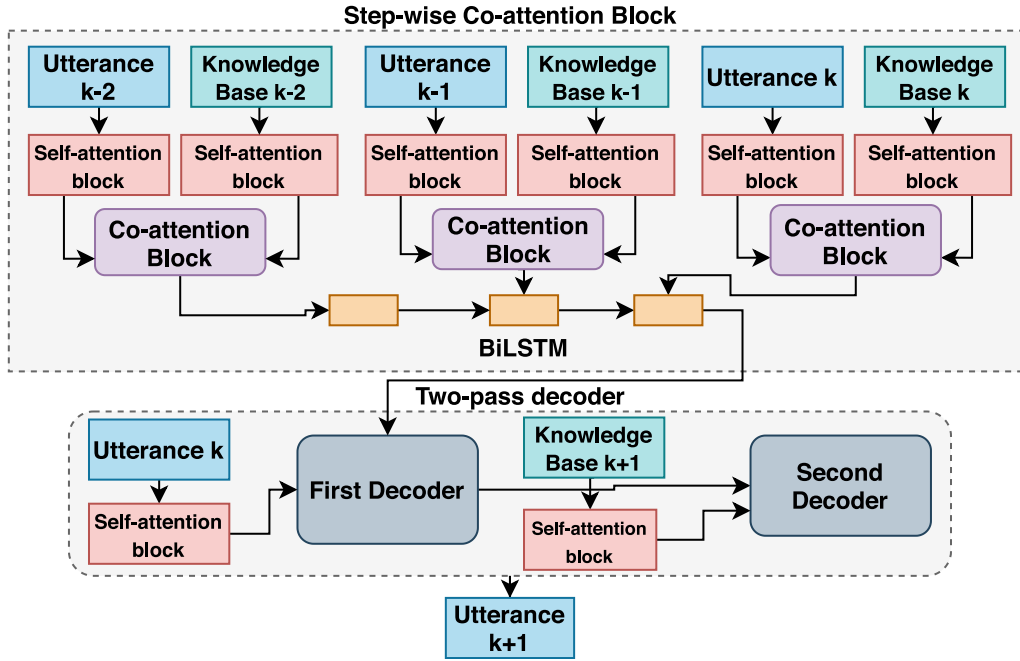


Fig. 1. Emo-STSC: Proposed model architecture. Here, the *Self-Attention Block*: is used to compute the representation of the knowledge base and utterance at every k th turn; The *Co-attention Block*: generates a representation of both the utterance and its corresponding knowledge base that is used by the *Dual Decoder* to decode responses. This approach allows the decoder to incorporate contextual and factual information from the knowledge bases.

3.1. Problem statement

We are given a sequence of K multi-turn utterances $U = u^{(1)}, \dots, u^{(k)}, \dots, u^{(K)}$, where $u^{(k)} = u_1^{(k)}, \dots, u_i^{(k)}, \dots, u_J^{(k)}$ represents the I words of the k th utterance. Corresponding to each utterance $u^{(k)}$, there is a knowledge base $d^{(k)} = d_1^{(k)}, \dots, d_j^{(k)}, \dots, d_J^{(k)}$ with J words, and the sequence of K knowledge bases is denoted by $D = d^{(1)}, \dots, d^{(k)}, \dots, d^{(K)}$. The sequence of labelled emotions associated with each utterance is represented as $E = e^{(1)}, \dots, e^{(k)}, \dots, e^{(K)}$. The objective is to produce a response $y = u^{(k+1)}$ given the set of previous k knowledge-grounded utterances, knowledge associated with gold response, $d^{(k+1)}$ and its emotion label $e^{(k+1)}$.

In this paper, we introduce an innovative solution for producing appropriate and relevant responses in a conversation using a transformer-based model with a Step-wise Co-attention (STSC) architecture and a binary classifier network. The architecture, as shown in Fig. 1, comprises three essential components: the Self-Attention Block (SA), Step-wise Co-attention Block (SC), and Dual Decoder Block (DD). For additional information, please consult Fig. 2.

- 1. Self-Attention Block:** To encode the user utterances and associated knowledge, the Self-Attention (SA) block uses a positional self-attention mechanism described in [52]. In Fig. 1, the self-attention blocks for the $(k-2)$ -th, $(k-1)$ -th, and (k) -th utterances and knowledge sentences are shown at the top.
- 2. Step-wise Co-attention Block:** We introduce a Step-wise Co-attention Block (SC) which receives the features of the current utterance $u^{(k)}$ and its corresponding knowledge base $d^{(k)}$ as inputs. The block employs a co-attention mechanism to generate $C_{U_o}^{(k)}$, an intertwined representation of the utterance and its associated knowledge base. The Co-attention Blocks are below the Self-attention block in Fig. 1 to integrate information from different sources.
- 3. Dual Decoder Block:** The architecture includes a transformer decoder based on multi-head self-attention, as explained in [8].

The first step of the decoder takes the self-attentive representation of the current utterance $u^{(k)}$ and the output of the step-wise co-attention block as input. This can be seen at the bottom of Fig. 1. The first step decoder relies on the context history for generating responses, while the second step decoder leverages the representation of the knowledge base, $d^{(k+1)}$, associated with the target utterance, $u^{(k+1)}$, to generate more accurate and relevant responses. This approach is inspired by human reasoning, where people often draft their responses based on the preceding statement in a conversation and use background information to refine their answers or ask questions. The dual decoder is designed to improve the coherence of response context and the accuracy of knowledge utilization.

3.2. Self-attention block

The most fundamental stage in the majority of natural language processing (NLP) activities is to translate words into numbers so that computers can recognize and decode linguistic patterns. This block utilizes a positional self-attention mechanism [52] to compute a representation of the current utterance and its related knowledge base. Let the input $I p_d$ represent the words from the knowledge base where the representation of each word is obtained by using the embedding, emb , of the current word, $emb(d_j)$, as well as the positional encoding $PE(j)$ of the j th word:

$$I p_d^{(k)} = [d_1^{(k)}, \dots, d_J^{(k)}] \quad (1)$$

$$d_j^{(k)} = emb(d_j^{(k)}) + PE(j) \quad (2)$$

The Self-Attention Block is a stack of N^x identical layers. There are two layers in the SA block. The first layer is a multi-head self-attention function followed by a position-wise fully connected feed-forward network (FFN).

$$S A_d(d^{(k)}) = FFN(\text{MultiHead}(I p_d^{(k)}, I p_d^{(k)}, I p_d^{(k)})) \quad (3)$$

where $n = 1, \dots, N^x$. x is the number of layers in the network. Similarly, we use $S A_u(u^{(k)})$ to denote the current utterance representation.

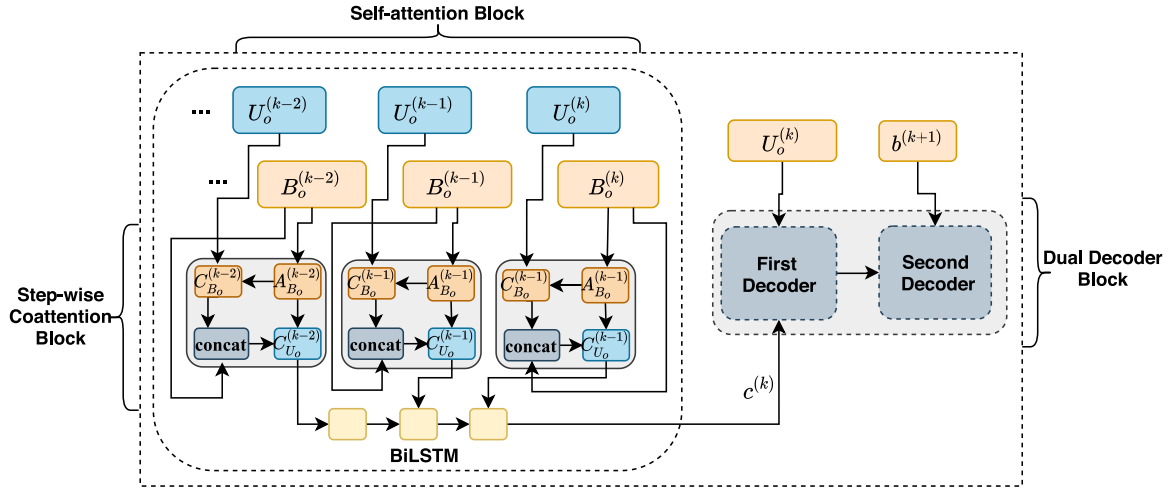


Fig. 2. Detailed architecture of the proposed model's components, as shown in Fig. 1. A simplified version of the Step-wise Co-attention block model is shown here to demonstrate the precision of our proposed STSC model. Our Step-wise Co-attention Block (SC) takes the current utterance $u^{(k)}$ and the corresponding knowledge base $d^{(k)}$'s features as inputs and then applies co-attention mechanism to compute $c_{U_o}^{(k)}$, an interdependent representation of the utterance and its related knowledge base.

3.3. Step-wise co-attention block

The generated response often does not need to reflect information from external resources such as Wikipedia, but it needs to combine the knowledge information properly on the basis of fully combining the historical dialogue information. To improve the knowledge performance of dialogue models, we exploit the latent implicit utterance and knowledge information as computed in Section 3.2 by a Step-wise Co-attention mechanism for dialogue context and knowledge information. This can be easily understood from the example shown in Table 1, the knowledge information of Agent 1 is seen in the utterance spoken by it. The co-attention module returns a vector that is attended by both the information from dialogue history and the most correlative knowledge information which is more fine-grained than existing models.

To build an effective multi-turn knowledge-grounded dialogue system, we utilize a Step-wise Co-attention [53] based encoder for encoding the context utterances as well as the associated knowledge incrementally. We use the positional self-attention block to obtain the attended features of utterance and knowledge: $B_o^{(k)} = SA_d(d^{(k)})$ and $U_o^{(k)} = SA_u(u^{(k)})$. Using the attended features, we compute the similarity scores corresponding to all the pairs of words from the knowledge base and utterance words: $S_o = (B_o^{(k)})^T U_o^{(k)}$

Bidirectional Attention Mechanism

We then compute the attention weights, $A_{U_o}^{(k)}$, across the knowledge base for every word in the context utterance by normalizing the similarity scores row-wise, and column-wise to get the attention weights, $A_{B_o}^{(k)}$, across the context utterances for every word in the knowledge base, i.e., $A_{U_o}^{(k)} = \text{softmax}(S_o)$ and $A_{B_o}^{(k)} = \text{softmax}(S_o^T)$

Co-attention: Next, we calculate the attention vectors of the context utterances with respect to each word of the knowledge base.

$$C_{B_o}^{(k)} = U_o^{(k)} A_{B_o}^{(k)} \quad (4)$$

Lastly, we compute $C_{U_o}^{(k)}$, a co-dependent representation of the utterance and knowledge base, as the co-attention context.

$$C_{U_o}^{(k)} = [B_o^{(k)}; C_{B_o}^{(k)}] A_{U_o}^{(k)} \quad (5)$$

where $[a; b]$ denotes the horizontal concatenation of the vectors a and b .

This step is repeated for all the k -utterances in the context history. We use $c^{(k)}$ to represent the encoded sentence. The final representation is obtained by passing all the previous k co-attended features through a hierarchical-encoder [4] model.

$$c^{(k)} = \text{BiLSTM}(C_{U_o}^{(k)}) \quad (6)$$

3.4. Dual decoder

During a conversation, humans often rely on their background knowledge about the topic to generate appropriate responses. To ensure a similar approach in our model, we integrate relevant knowledge information during the decoding stage to improve response diversity. This is done by utilizing the dialogue context to generate a response and incorporating the most correlated factual knowledge to assist in the response generation process.

To decode the responses, we adopt a dual decoder approach that sequentially processes the encoded context. The first decoder consists of four sub-parts, the initial being multi-head self-attention, followed by multi-head context attention in the second part. The third part involves multi-head attention on the last utterance, while the fourth part is a position-wise fully connected feed-forward network, just like before.

$$I p_r^{k+1} = u_{\leq i}^{(k+1)} = \{u_0^{k+1}, u_1^{k+1}, \dots, u_{i-1}^{k+1}\} \quad (7)$$

$$E_1^n = \text{MultiHead}(I p_r^{k+1}, I p_r^{k+1}, I p_r^{k+1}) \quad (8)$$

$$F_1^n = \text{MultiHead}(E_1^n, c^{(k)}, c^{(k)}) \quad (9)$$

$$G_1^n = \text{MultiHead}(F_1^n, SA_u(u^{(k)}), SA_u(u^{(k)})) \quad (10)$$

$$H_1^n = FFN(G_1^n) \quad (11)$$

$$P(\hat{y}_{(1)}) = P(\hat{u}_{(1)}^{(k+1)}) = \text{softmax}(H_1^n) \quad (12)$$

Initially, the second decoder of our architecture includes a layer of multi-head self-attention, which is followed by a decoder layer for multi-head knowledge attention, and finally, there is another decoder layer for multi-head attention on the outputs of the first decoder. The last sub-layer of the second decoder is a position-wise fully connected feed-forward network.

$$E_2^n = \text{MultiHead}(I p_r^{k+1}, I p_r^{k+1}, I p_r^{k+1}) \quad (13)$$

$$F_2^n = \text{MultiHead}(E_2^n, SA_d(d^{(k+1)}), SA_d(d^{(k+1)})) \quad (14)$$

$$G_2^n = \text{MultiHead}(F_2^n, SA_u(\hat{u}_{(1)}^{(k+1)}), SA_u(\hat{u}_{(1)}^{(k+1)})) \quad (15)$$

$$H_2^n = FFN(G_2^n) \quad (16)$$

$$P(\hat{y}_{(2)}) = P(\hat{u}_{(2)}^{(k+1)}) = \text{softmax}(H_2^n) \quad (17)$$

We use a binary classifier to determine the emotion associated with each utterance, with the classifier's output based on the emotion labels. To accomplish this, we combine the predicted response $\hat{u}^{(k+1)}$ with its correct emotion label $e^{(k+1)}$ in a joint hidden representation and feed it as a single input to the Multi-layer Perceptron (MLP) hidden layer.

The binary classifier is trained on both true and fake examples, with true examples being those with correct emotion categories and fake examples being those with incorrect emotion categories. To generate the necessary samples, we used the topical chat and ISEAR datasets and set a threshold of 3070. We added additional instances from the ISEAR dataset to balance the number of instances in the disgusted, angry, and fearful classes, as they had fewer samples than the threshold value. We also added fake samples by randomly assigning non-neutral emotion classes to the remaining samples from the “Neutral” emotion class of the topical chat dataset. We collected a total of 22,792 fake samples, equal to the number of true samples. Table 2 displays the frequency of emotion classes in the topical chat dataset.

$$D(y/U, E) = \sigma(\text{MLP}([\hat{u}^{(k+1)}, e^{(k+1)}])) \quad (18)$$

The notation $[\cdot, \cdot]$ means combining or joining two elements together, while σ refers to a specific mathematical function called the Sigmoid function.

Training: We aim to minimize the following loss similar to [8]:

$$Loss = L_{mle_1} + L_{mle_2} + L_{classifier} \quad (19)$$

$$L_{mle_1} = - \sum_{i=1}^I \log P(\hat{y}_{(1)i}) \quad (20)$$

$$L_{mle_2} = - \sum_{i=1}^I \log P(\hat{y}_{(2)i}) \quad (21)$$

$$L_{classifier} = -\log(D(y/U, E)) - \log(1 - D(P(\hat{y}_{(2)}))) \quad (22)$$

4. Datasets and experimental setup

In this section, we present the datasets, baselines, and experimental results, along with the necessary analysis.

4.1. Dataset

We evaluate our model on two benchmark datasets, the Topical Chat dataset [10] and Document Grounded Conversations dataset [9].

4.1.1. Topical chat dataset

The dataset is grounded on external knowledge which is composed of multiple sentences for evaluating our proposed model. The dataset consists of approximately 11,000 conversations between two humans on topics such as politics, sports, fashion, books, music, general entertainment, science and technology, and movies. Some of the utterances were written by the annotators without any accompanying knowledge. The dataset includes information on the emotion associated with each utterance. The most common emotion labels are sad, surprised, angry, disgusted, fearful, curious to dive deeper, happy, and neutral. The data is divided into five categories: train, valid frequently, valid rarely, test frequently, and test rarely. The Frequent and Rare sets contain entities (i.e., the topic of the conversation (Fashion, Politics, etc.) that were frequently and rarely seen in the training set, respectively. Table 3 provides the size details of the training, test, and validation dataset. The frequency of every emotion class is shown in Table 2.

4.1.2. Document grounded conversations dataset (CMU_DoG)

There are 72,922 utterances in the training set, 3,626 utterances in the validation, and 11,577 utterances in the testing set. The utterances are grounded on related topics like movie names, casts, introductions, ratings, and some scenes. The documents have an average length of around 200.

We utilize an emotion classifier based on BERT trained on the Topical Chat dataset utterances and the target utterances. Our model’s performance is evaluated on 200 sentences from the test set, and we attain an accuracy score of 0.78 overall. The dataset information is presented in Table 3.

Table 2

Different emotion classes and their frequency distribution in the topical chat dataset.

Emotion classes	Original count
Curious to dive deeper	101 162
Happy	36 845
Sad	3070
Angry	1133
Surprised	38 254
Neutral	51 796
Disgusted	1848
Fearful	1174

It is important to note that a significant portion of the knowledge incorporated into our study originates from open-source Wikipedia pages, which are publicly available online. These Wikipedia documents serve as a valuable source of information, offering a diverse range of topics and ensuring the reliability and accuracy of the knowledge integrated into our system.

4.2. Baselines

We benchmark our model against various baselines, focusing on emotion-based, knowledge-based, and integrated emotion-knowledge dialogue tasks.

4.2.1. Knowledge grounded dialogue generation

- Know-Seq2Seq:** We first develop basic knowledge-grounded baseline models by extending traditional encoder–decoder architecture [3] with dot product attention [54] and feeding the dialogue utterances concatenated with the grounded knowledge as the input to the model.
- Know-Transformer:** Our second baseline model is based on a multi-head attention mechanism [52] which was originally proposed for Neural Machine Translation (NMT) systems. The input to the model is again the concatenation of utterances and the corresponding knowledge base.
- Know-HRED:** Our third baseline is the extension of the hierarchical recurrent encoder–decoder model (HRED). In addition to the Seq2Seq model, it employs a context RNN which iteratively processes all the sentence vectors of a conversation and updates its hidden state after every sentence [55]. Every sentence vector is a representation of context utterance along with the grounded knowledge base.
- ITDD:** [8] Our subsequent baseline utilizes the generative Incremental Transformer with a Deliberation Decoder rooted in the transformer framework. This model incrementally constructs the representation of context utterances and knowledge sentences using a multi-head attention mechanism. The decoder has two stages: the initial focuses primarily on the conversational context, while the latter incorporates document knowledge to enhance the response generation.
- GPT-2_{finetune}** [56]: We use the OpenAI GPT-2 with 345 parameters as our fifth baseline model. We first concatenate all dialogue turns inside a dialogue into a long text, terminated by the end-of-text token, for each dialogue sample in the Topical Chat and CMU_DoG dataset, and then we optimize the model with the language modelling objective.
- ZRKG** [36]: A double latent variable model is proposed that captures both the knowledge relating a context and response as well as the manner in which the knowledge is expressed. It also suggested a variational learning strategy to use outside knowledge sources as response generators.

Table 3
Dataset details.

Topical chat						CMU_DoG		
	Train	Valid frequent	Valid rare	Test frequent	Test rare	Train	Valid	Test
#Conversation	8628	539	539	539	539	3,373	229	619
#Utterances	188 378	11 681	11 692	11 760	11 770	74,717	4,993	13,646
Avg. # of Turns	21.8	21.6	21.7	21.8	21.8	22.2	21.8	22.0

7. **KnowledgeGPT**: [33] The KnowledgeGPT model employed a knowledge selection module, which reduces the redundant knowledge input with pertinent data by formalizing knowledge selection as a sequence prediction process, and an autoregressive GPT-2 model with policy gradient optimization networks.
8. **DiffKS**: [32] They suggest a strategy for selecting knowledge that is difference-aware. The difference between the candidate knowledge sentences offered during the current turn and those selected during the preceding turns is computed first. The differential information is then combined with or separated from the contextual information to aid in the choice of ultimate knowledge.
9. **KAT** [35]: They proposed a knowledge-aware Transformer with a dynamic knowledge selection mechanism, which can completely utilize the external knowledge to produce fluent and insightful dialogue responses.
10. **STSC**: This specific model setup is a component of our Emo-STSC model, which exclusively utilizes the external knowledge connected to the utterances. It includes a self-attention section that encodes the knowledge sentences and utterances and utilizes the Co-attention Block to create a mutually dependent representation of the utterance and its connected knowledge base. Finally, it uses a decoder based on transformers to produce the responses.

4.2.2. Emotion grounded dialogue generation

1. **Emotional Chatting Machine** [37]: Emotional grounding in dialogue models is captured using this model. It incorporates three mechanisms emotion category embedding, internal emotion memory, and external memory into the existing seq2seq model.
2. **CAiRE** [46]: The Generative Pre-trained Transformer (GPT) in this system is modified using transfer learning to address an empathetic response generating challenge. It makes use of the GPT-2 model that has been previously trained in unsupervised settings on the BooksCorpus dataset.
3. **Emo-Transformer**: We developed the Emo-Transformer to encode the input utterances and solely use the emotion labels connected with them in order to better understand the function of our emotion discriminator, which is conditioned on the emotion labels associated with each utterance.

4.2.3. Emotion-and-knowledge grounded dialogue generation

1. **EmoK-Seq2Seq**: For both emotion and knowledge-grounding, we begin by employing an emotion classifier to the traditional encoder-decoder architecture with dot product attention [54] in order to generate emotionally relevant responses. However, for knowledge grounding, we follow a similar architecture as Know-Seq2Seq and give as input the concatenated utterance and knowledge sentences.
2. **EmoK-Transformer**: Similarly, our second baseline is prepared by augmenting an emotion-specific classifier at the end of a multi-head attention-based transformer model to generate a response based on the emotion labels associated with each utterance. The input to the model however remains the same as before, merged utterance and knowledge sentences for knowledge-grounded generation.

3. **EmoK-HRED**: The HRED [4] model is extended to take both the knowledge and context utterances as input to engage the external knowledge for dialogue generation. It is further extended to include emotion labels also as input for training an end-to-end neural knowledge and emotion-grounded dialogue model.
4. **EmoK-ITDD**: We further extended the state-of-the-art KG-based incremental transformer ITDD [8] model to incorporate a classifier in order to produce responses based on the emotion labels associated with each utterance ultimately creating a baseline for knowledge and emotion grounded dialogue system.
5. **EmoKbGAN**: We adopt the cutting-edge EmoKbGAN model proposed by [7] as our baseline. This model leverages a GAN-based architecture to generate dialogues that are grounded in both knowledge and emotion. Instead of using the traditional Maximum Likelihood Estimation (MLE) objective, EmoKbGAN utilizes multi-attribute discriminator training to supervise the training process.

4.3. Experimental setup

The input to the model is the previous three utterances termed the context history and their related text-based knowledge base. We use Glove word embedding [57] for representing the input sequences. For implementation, we use OpenNMT-py¹ [58]. We set the hidden dimension to 512 for all the models. For our Seq2Seq and transformer-based models, we use a 3-layer encoder and decoder. In multi-head attention, the number of attention heads is 8, and the size of the hidden transformer feed-forward is 2,048. We use the ADAM optimizer [59] with a learning rate of 0.0001 and beam search with a beam width equal to 5 for generating responses. For replicating the state-of-the-art model, we use the code provided by [8].²

4.4. Evaluation metrics

4.4.1. Automatic evaluation

In order to evaluate our generated response on both frequent and rare test sets, we use the standard metrics like BLEU [60], F1-score, perplexity (PPL) and n-gram diversity (Div.) [6]. Perplexity (PPL) is a measure used to evaluate the effectiveness of a probability model in predicting a sentence. Typically, a lower PPL suggests superior predictive performance. The BLEU score compares consecutive phrases of words in the generated and gold response. N-gram diversity is used as a measure of informativeness and diversity of sentences. We also compute the *unigram* F1-score, which is the difference between the model prediction and the ground truth responses. We use the validation set to report the perplexity scores. All the above metrics compare the predicted responses with the gold responses. Although it seems relevant to use human evaluation results for model comparisons as in open-domain settings one can have an adequate response which is completely different from the ground truth response. We still provide automatic evaluation results for a better understanding of the data. By assessing the emotional content of the predicted words using an accuracy metric, we may assess how well the emotion classifier performs.

¹ <https://github.com/OpenNMT/OpenNMT-py>

² <https://github.com/lizekang/ITDD>

Table 4

Experimental results using automatic and manual evaluation metrics for all the models as defined in Section 4.2 on the Topical Chat dataset. The leading results for each metric are indicated by a boldface.

Models	PPL (Freq)	BLEU% (Freq/Rare)	F1 (Freq/Rare)	Div.(n=1) (Freq/Rare)	Div.(n=2) (Freq/Rare)	Fluency (Freq)	Adequacy (Freq)	Emotional content (Freq)	Knowledge relevance (Freq)
Knowledge-Grounded Dialogue Generation									
Know-Seq2Seq	82.77	0.95/0.74	0.20/0.17	0.020/0.034	0.044/0.075	1.22	0.35	0.25	0.22
Know-Transformer	150.01	0.37/0.25	0.16/0.16	0.013/ 0.013	0.032/ 0.031	0.82	0.38	0.18	0.22
Know-HRED	82.77	0.59/0.58	0.23/0.20	0.025/ 0.026	0.042/ 0.053	1.48	0.38	0.28	0.22
ITDD [8]	55.29	1.04/0.88	0.23/0.19	0.082/ 0.112	0.231/0.300	1.59	0.71	0.38	0.48
<i>GPT-2_{finetune}</i> [56]	13.40	1.10/0.80	0.13/0.13	0.030/ 0.024	0.043/0.068	1.50	0.41	0.12	0.22
DiffKS [32]	33.40	1.08/0.78	0.15/0.14	0.069/ 0.079	0.107/0.153	1.68	0.54	0.16	0.27
ZRKG [36]	25.91	1.05/0.75	0.15/0.15	0.075/ 0.080	0.102/0.148	1.63	0.56	0.19	0.36
KAT [35]	23.40	1.10/0.76	0.15/0.15	0.079/ 0.076	0.104/0.168	1.67	0.59	0.30	0.45
KnowledGPT [33]	21.42	1.12/0.80	0.16/0.16	0.085/ 0.115	0.234/0.302	1.62	0.64	0.31	0.41
STSC	47.88	1.14/0.84	0.25/0.22	0.044/ 0.054	0.101/0.118	1.76	0.79	0.40	0.55
Emotion-Grounded Dialogue Generation									
ECM [37]	89.43	0.45/0.32	0.10/0.09	0.027/ 0.036	0.058/0.079	1.43	0.38	0.25	0.22
CAiRE [46]	74.63	0.53/0.41	0.11/0.10	0.024/ 0.034	0.053/0.072	1.57	0.42	0.29	0.23
Emo-Transformer	63.58	0.58/0.44	0.15/0.14	0.038/0.037	0.056/0.049	1.50	0.49	0.35	0.21
Emotion and Knowledge Grounded Dialogue Generation									
EmoK-Seq2Seq	75.68	0.95/0.92	0.20/0.18	0.023/ 0.035	0.055/0.076	1.25	0.38	0.22	0.28
EmoK-Transformer	133.58	0.35/0.20	0.17/0.16	0.008/0.007	0.022/0.019	0.90	0.39	0.25	0.30
EmoK-HRED	80.06	0.85/0.68	0.23/0.20	0.016/ 0.018	0.033/0.031	1.08	0.25	0.22	0.20
EmoK-ITDD	47.90	1.34/1.12	0.23/0.21	0.079/ 0.113	0.217/ 0.305	1.66	0.74	0.41	0.55
EmoKbGAN	88.8	0.91/0.70	0.20/0.18	0.076/ 0.091	0.189/0.181	1.87	1.17	0.52	0.55
Emo-STSC	57.10	1.15/0.80	0.25/0.22	0.084/0.106	0.239 / 0.292	1.89	1.20	0.53	0.57
STSC-EC	47.88	1.14/0.84	0.25/0.22	0.044/ 0.054	0.101/0.118	1.76	0.79	0.40	0.55
STSC-DD	42.80	0.92/0.65	0.20/0.18	0.034/ 0.045	0.096/0.102	1.67	0.67	0.35	0.49
STSC-coattnB	36.73	0.68/0.43	0.15/0.14	0.029/ 0.042	0.093/0.101	1.54	0.56	0.32	0.43

As a result, in order to evaluate the performance of our generation model, we additionally compute the emotional correctness of the generated answer using our pre-trained classifier. By leveraging the encoded context knowledge information to choose pertinent knowledge phrases, we also employ accuracy as a criterion to assess the relevance of the knowledge in the predicted response.

4.4.2. Human evaluation

To evaluate our proposed model, we generate 20 random sample conversations which contain a total of 436 utterances from the frequent test set of the Topical chat dataset for evaluation. For the Document Grounded Dataset, we again randomly sampled around 30 conversations containing 384 utterances for evaluation purposes. To ensure robustness in this assessment, we engaged two seasoned professionals equipped with post-graduate degrees to function as our human judgment annotators. These annotators are not only integral members of our research team, but they also possess an in-depth familiarity with projects akin to the present study, having been associated with them over the past three years. Both are employed on a regular basis in accordance with university norms, drawing a monthly remuneration of Rs 35,000. For each test case, the annotators were presented with model-generated responses juxtaposed with their human-constructed counterparts to facilitate a comprehensive evaluation. We evaluate the responses using the following metrics:

(i) Fluency: It is a property of a sentence through which we measure its grammatical complexity. (ii) Adequacy: It measures whether the generated response is in accordance with the context. (iii) Knowledge Relevance [5]: It checks whether the generated response is relevant in view of the associated knowledge base. (iv) Emotional Content:

Emotional content refers to the extent to which the generated response conveys the intended or target emotion.

The scores for fluency, adequacy, and knowledge relevance range from 0 to 2, with 0 indicating an incomplete or incorrect sentence, 1 indicating a decent sentence, and 2 indicating a perfectly correct sentence. The emotional content is measured on a scale of 0 to 1, where 0 represents an incorrect emotion and 1 represents a correct emotion. To assess the agreement between two annotators, the Fleiss' kappa score [61] is calculated. The kappa scores for emotional content, knowledge relevance, fluency, and adequacy for the Topical chat dataset are 0.81, 0.72, 0.87, and 0.76, respectively, indicating "high agreement". For the document-grounded conversation dataset, the kappa scores for emotional content, knowledge relevance, fluency, and adequacy are 0.78, 0.86, 0.71, and 0.68, respectively, also denoting "high agreement".

5. Results and analysis

We have structured our experiments to address the following key questions:

1. **Performance Comparison:** How does the proposed method perform compared to state-of-the-art methods? This question will be thoroughly explored in automatic evaluations (Section 5.1), Ablation study (Section 5.5), and human evaluations (Section 5.2). Table 4 and Table 5 show the performance of various models on the test set of the two datasets, respectively. We provide a discussion on comparison with the state of the art in Section 5.6.

Table 5

On the Document Grounded Conversation dataset, experimental results were obtained using automatic and manual evaluation metrics for all models defined in Section 4.2. The boldface indicates the best results for each metric. The bolded result is after 30 epochs.

Models	PPL	BLEU%	F1	Div.(n=1)	Div.(n=2)	Fluency	Adequacy	Emotional content	Knowledge relevance
Knowledge-Grounded Dialogue Generation									
Know-Seq2Seq	80.83	0.35	0.12	0.022	0.049	1.66	0.51	0.20	0.14
Know-Transformer	88.32	0.39	0.13	0.018	0.039	1.58	0.63	0.21	0.27
Know-HRED	81.14	0.44	0.15	0.028	0.052	1.30	0.31	0.20	0.19
ITDD [8]	37.06	1.08	0.18	0.039	0.135	1.68	0.92	0.30	0.54
<i>GPT-2</i> _{finetune} [56]	17.83	0.60	0.09	0.021	0.055	1.38	0.33	0.18	0.21
DiffKS [32]	45.50	0.84	0.09	0.059	0.134	1.65	0.58	0.21	0.32
ZRKGK [36]	54.50	0.95	0.10	0.045	0.129	1.56	0.50	0.18	0.25
KAT [35]	22.53	0.96	0.13	0.059	0.134	1.63	0.54	0.19	0.29
KnowledGPT [33]	20.53	0.98	0.15	0.064	0.145	1.67	0.59	0.20	0.33
STSC	37.88	1.06	0.19	0.064	0.234	1.76	1.04	0.34	0.60
Emotion-Grounded Dialogue Generation									
ECM [37]	78.13	0.40	0.14	0.026	0.045	1.35	0.32	0.16	0.23
CAiRE [46]	64.24	0.53	0.15	0.035	0.065	1.43	0.38	0.17	0.19
Emo-Transformer	60.35	0.61	0.15	0.039	0.062	1.52	0.42	0.19	0.25
Emotion and Knowledge Grounded Dialogue Generation									
EmoK-Seq2Seq	77.67	0.38	0.13	0.024	0.055	1.50	0.61	0.24	0.22
EmoK-Transformer	70.35	0.66	0.15	0.019	0.032	1.62	0.56	0.18	0.35
EmoK-HRED	79.12	0.77	0.17	0.030	0.057	1.56	0.47	0.20	0.35
EmoK-ITDD	33.39	1.07	0.18	0.055	0.203	1.66	0.94	0.48	0.55
EmoKbGAN	38.50	1.10	0.16	0.065	0.156	1.80	1.25	0.50	0.60
Emo-STSC	38.81	1.10	0.20	0.078	0.231	1.82	1.27	0.68	0.63
STSC-EC	37.88	1.06	0.19	0.064	0.234	1.76	1.04	0.34	0.60
STSC-DD	30.32	0.87	0.15	0.046	0.178	1.67	0.87	0.30	0.54
STSC-coattnB	37.88	1.06	0.19	0.064	0.234	1.54	0.7	0.34	0.60

- Statistical Significance:** To ensure the reliability of our results, we delve into the statistical significance of our findings in Section 5.3.
- Case Studies:** To provide deeper insights into the real-world applicability and nuances of our method, we present case studies in Section 5.4, illustrating the performance of the proposed methods.

5.1. Automatic evaluation results

As our experimental setup is primarily divided into three sections, KGDG (Knowledge-Grounded Dialogue Generation), EGDG (Emotion-Grounded Dialogue Generation), and EKGDG (Knowledge-and-Emotion Grounded Dialogue Generation), we carry out our study separately for each section. Table 4 reports evaluation results on Test Freq and Test Rare of Topical Chat, and Table 5 reports evaluation results on CMU_DoG. For the knowledge-dependent study, in Table 4, we follow from the n-gram diversity scores that the proposed STSC model for the sub-task of KGDG, despite being inferior to several of the baseline models, has high unigram and bigram diversities on both the frequent and rare test set, demonstrating that the decoded responses have a high level of diversity and informativeness. In particular, when compared to other state-of-the-art models for the task of KGDG, STSC, which solely uses grounded knowledge, outperforms *ITDD*, *GPT-2*, *ZRKGK*, *KnowledGPT*, *DiffKS*, and *KAT*.

We observe from Table 4 that on the topical chat dataset, KnowledGPT obtains the highest F1 scores of 0.16/0.16 and BLEU score of 1.12/0.80 on the freq and rare test set of Topical chat dataset among the baselines, whereas the STSC model outperforms the best baselines with a BLEU and F1 score of 1.14/0.84 and 0.25/0.22, respectively. This observation suggests that the co-attention mechanism offers a simple yet effective way to improve the encoding of the knowledge base and context utterances in order to produce more engaging responses. This is in contrast to the KnowledGPT model, which uses knowledge selection as a technique to select relevant knowledge that may not always be correct and thereby lead to erroneous responses.

For the sub-task of emotion-grounded dialogue generation using the Topical Chat dataset, we may again observe in Table 4, that our proposed model with only the emotion module and without any external knowledge, i.e., Emo-Transformer model outperforms the state of the art, ECM and CAiRE models on Diversity, F1 and BLEU scores. It is interesting to note that even though the best baseline viz. the CAiRE model uses a large pre-trained language model, learning from a discriminator produces better performance. All the models with the prefix ‘‘Emo’’ utilize the class labels as emotions and tend to perform better with higher BLEU and F1 scores. We can infer from the results above that the co-attention-based response generation model can be guided by learning from binary classifiers to generate relevant replies in accordance with the target emotions.

The joint training model viz. the knowledge-and-emotion grounded model (*Emo-STSC*), outperforms the results in both separate sub-tasks on the Topical Chat dataset viz. KGDG and EGDG. This experiment is conducted to study the effect of guided learning by leveraging information from multiple sources (i.e., emotion and knowledge). It can be observed that the results for all the metrics improve, except perplexity. This is due to the fact that PPL is determined using real data from test sets, making models learned by fitting the same or a related distribution more favourable on the measure. Also, since many of the models, such as GPT-2 are models pre-trained on large data, it limits the model’s performance and prevents it from picking up contextual understanding, which lowers the baseline models’ BLEU and F1 scores as shown in Table 4.

Additionally, we run tests to see how well the suggested approach performs with datasets that have many fewer sample conversations. We perform experiments on the CMU_DoG having 3K conversations which is 1/3rd of the Topical chat dataset. We observed that our proposed Emo-STSC performs better on the Topical Chat dataset than Document Grounded viz. CMU_DoG dataset. From Table 5, we observed that our proposed STSC model obtains the highest BLEU and F1 scores for the knowledge-grounded subtask. We also observe for the emotion grounded subtask viz. EGDG, our proposed model (*Emo-Transformer*) outperforms the baseline, ECM and CAiRE models. Since, in CMU_DoG, the crowd-workers do not refer to external knowledge as much as those workers do in Topical Chat when they form the responses.

Hence, the joint learning of emotion and knowledge, i.e., the *EKGDG* experiments results in outperforming the state-of-the-art scores on the individual tasks. The co-attention step and dual decoding step are advantageous because they enable the fusion of knowledge into conversational context information, which improves the use of representations obtained from transformer models. Moreover, by using the dual decoding step, we can benefit from the knowledge output gradually. The ablation models' lower perplexity ratings provide additional evidence of their capacity to produce better responses.

We obtain an emotional classification accuracy of 0.78 and 0.73 on the Topical Chat and CMU_DoG datasets, respectively for our proposed approach. We also obtain an accuracy of 0.69 and 0.73 for knowledge selection by our encoder model as described in Sections 3.2 and 3.3 on the test set of both Topical Chat and CMU_DoG dataset.

5.2. Human evaluation results

For determining the quality of the responses, human evaluation is just as important as automatic evaluation. Since human evaluation is expensive, we only assess the Topical Chat dataset's frequent test. We compare our baseline and suggested approach using the human assessment metrics for the *viz.* tasks (KGDG, EGDG, and EKGDG). In Tables 4 and 5, we present the results of the human evaluation for all the tasks, baselines, and the proposed model for both datasets. An important statistic to assess the responses is fluency, which ensures that the grammar is correct. Due to the Know-Transformer model's use of the straightforward fusion approach of concatenation in this framework, it generates the lowest fluency scores for the task of knowledge-based conversation generation. The models such as DiffKS and KAT, show high scores for fluency denoting efficient fusion methods used in the frameworks. However, as seen in Tables 4 and 5, our proposed STSC framework outperforms the baseline models by achieving the highest scores for fluent responses with an increase of approximately 8% on both datasets.

Similar to this, the suggested framework for the adequacy metric indicates an improvement of roughly 2 points over the baseline techniques. This may indicate that the suggested framework produces contextually appropriate responses. The generated responses are evaluated in accordance with the designated modalities because the focus of the current work is on knowledge and emotion. The results, which are displayed in Tables 4 and 5, clearly demonstrate that the suggested framework produces responses that are relevant to the stated modalities with a *Knowledge Relevance* score of 0.55 and 0.60, respectively, for both the dataset. The results presented in Tables 4 and 5 demonstrate that the proposed STSC framework outperforms the baseline models in terms of fluency scores by approximately 8% on both datasets. Additionally, the emotional metric shows an improvement of about 1 to 2 points compared to the KnowdGPT and ITDD baselines on both the Topical Chat and CMU_DoG datasets. These findings suggest that incorporating knowledge information in the frameworks leads to better human evaluation performance.

In the case of Emotion-Grounded Dialogue Generation, it is evident from the results in Tables 4 and 5 that discriminator-based approaches outperform the traditional frameworks in terms of human evaluation also. This is primarily due to the networks' improved guiding abilities, which can produce responses that are consistent with the attribute. By revealing more about the speaker's preferences and mood, the emotional information enhances performance by limiting the generation of responses to the emotion being discussed. As a result, according to the Emotional Content metric, our proposed framework with the emotion information and classifier network outperforms baseline approaches like ECM and CAiRE by 1%–2% on both datasets. The generated responses are not only fluent and pertinent but also congruent with the emotion description, according to the human evaluation.

For the third task of Emotion-and-Knowledge Grounded Generation, in terms of all the human evaluation metric scores on both the Topical Chat and CMU_DoG dataset, our proposed model Emo-STSC demonstrates superior performance when compared to the other baseline models, as illustrated in Tables 4 and 5. On observing the manual evaluation scores, we find that most of the responses from our proposed model (Emo-STSC) are readable and highly coherent with a score of 1.89 and 1.20 on the Topical Chat dataset and a score of 1.80 and 1.25 on the CMU_DoG dataset when they exploit both the available knowledge and emotion attributes. We ensure that our responses are both fluent and understandable by comparing the fluency and adequacy scores of our model to those of the baseline models. Additionally, we evaluate the emotional content score and knowledge relevance score to confirm that our generated responses are both emotionally appropriate and accurately grounded in the given knowledge. On the other hand, our Emo-ITDD and STSC-EC models have also achieved good scores on all four measures, indicating an overall performance improvement by using the respective modules (co-attention and emotion classifier) in the architectures with respect to the ITDD model.

5.3. Statistical significance test

For the KGDG task, the improvement in scores of the STSC model is statistically significant when compared with the best baseline model, i.e., KnowledGPT and ITDD models for the BLEU, F1 scores and human evaluation metric scores on the Topical Chat dataset as shown in Table 4, and is also statistically significant when compared to the best baseline model, i.e., KnowledGPT and ITDD models for the BLEU, F1, Diversity and human evaluation metric scores on the CMU_DoG dataset as shown in Table 5. We test this by performing statistical significance tests (t-test) [62], and it is conducted at 5% (0.05) significance level.

For the EGDG task, we perform a statistical significance test (T-test) on the obtained results and observe that performance improvement in the proposed model over the state of the art, ECM and CAiRE model on all the metrics of automatic and human evaluation is significant with 95% confidence (i.e. p -value ≤ 0.05) on both the datasets, Topical Chat and CMU_DoG.

Significance tests were conducted between the best results and other competitors on the EKGDG tasks also, with t-test for F1, Diversity on a freq dataset of Topical Chat; BLEU, F1, Div(n=1) on CMU_DoG and human evaluation scores on both the dataset. The experiment was statistically significant as they indicated p -value $\leq 0.05/0.005$.

5.4. Detailed analysis

We perform a detailed qualitative analysis of the responses to further analyse the outputs generated from our proposed model. We show predictions from the state-of-the-art systems from each task such as ITDD, and KnowledGPT, for KGDG; CAiRE from EGDG and Emo-ITDD, and Emo-STSC from EKGDG. In Table 6, we observe in the first example that, Emo-STSC generates an adequate as well as an emotionally relevant response, whereas Emo-ITDD generates responses that are fluent but is inconsistent with the associated facts. In the example of Table 7, we observe factual as well as emotional inconsistency by the Emo-ITDD model. The performance improvement by our proposed approach might be attributed to the bidirectional attention mechanism that is based on a shared similarity matrix and operates in two directions, *viz.* knowledge-to-utterance as well as utterance-to-knowledge, and efficiently combines the utterance and knowledge to create a feature vector that generates responses that are both relevant in context and emotional. Along with a binary classifier that aids in the model's generation of an informed and emotion-controlled response, the second decoder also effectively learns from the knowledge base during decoding. Similar behaviour was observed for the CMU_DoG dataset as shown in Tables 8 and 9.

Table 6

Some sample responses generated by the baseline and proposed models on our test set from the Topical chat dataset. The sentence used from the corresponding knowledge base is highlighted in bold.

Utterance 1	do you like football?	Knowledge Base: Football is a family of team sports that involve, to varying degrees, kicking a ball to score a goal. Unqualified, the word football is understood to refer to whichever form of football is the most popular in the regional context in which the word appears. Sports commonly called football in certain places include association football (known as soccer in some countries); gridiron football (specifically American football or Canadian football); Australian rules football; rugby football (either rugby league or rugby union); and Gaelic football. These different variations of football are known as football codes. the University of Iowa's locker room for visiting football teams is completely painted pink. In the 1960's top bowlers made twice as much as top football stars - That the highest score ever in a football game occurred in 1916 when Georgia Tech defeated Cumberland 222-0 ESPN won an emmy for the creation of the superimposed yellow line representing the first down line for American football games. Former Patriots RB BenJarvus Green-Ellis has never fumbled the football in his NFL career.
Utterance 2	I sure do.. Did you know U of Iowa painted the visitor locker room pink?	
Utterance 3	I wonder if it is supposed to have some effect on them.	
Utterance 4	I wonder too. Did you know bowlers made more than football players in the 60s?	
(Gold)	Emotion: Curious to Dive deeper	
Emo-STSC	I wonder too. Did you know bowlers made more than football players? Emotion: Curious to Dive deeper	
Emo-STSC-F	Good point. Did you know bowlers made more than football players ? Emotion: Curious to Dive deeper	
Emo-ITDD	Good point. Do you know who BenJarvus Green-Ellis is? Emotion: Curious to Dive deeper	
ITDD	Yeah. Do you know who BenJarvus Green-Ellis is? Emotion: Curious to Dive deeper	
KnowledGPT	Yeah. I know the University of Iowa's locker room for football teams is completely painted pink. Emotion: Neutral	
STSC	Good point. Did you know bowlers made more than football players? Emotion: Curious to Dive deeper.	
CAiRE	I love watching football. Emotion: Happy	
Emo-Transformer	I wonder if they painted the locker room pink. Emotion: Curious to Dive deeper	
EmoKbGAN	university of iowa's locker room is painted pink. Emotion: Neutral	
ChatGPT	That's interesting. Football is a fascinating sport with different variations depending on the regional context. The fact about the University of Iowa painting the visitor locker room pink is intriguing; I wonder if there's any significance or effect on the visiting teams. Speaking of football facts, did you know that top bowlers used to earn more than top football stars in the 1960s? Another surprising fact is that the highest score ever in a football game was in 1916, where Georgia Tech defeated Cumberland 222-0. Also, ESPN won an Emmy for creating the superimposed yellow line that represents the first down line for American football games. Another fun fact is that BenJarvus Green-Ellis, a former Patriots running back, has never fumbled the football in his NFL career. Emotion: Curious to Dive deeper	

Table 7

Some sample responses generated by the baseline and proposed models on our test set from Topical chat dataset. Sentence used from the corresponding knowledge base is highlighted in bold.

Utterance 1	I get that, especially when there is only 11 min of action in a game. I guess the rest is commercials?	Knowledge Base: The NFL has no written rule against female players; women would in fact be allowed if they met the league's eligibility requirements. an average NFL game only has 11 min of live gameplay The Average Lifetime Earnings of an Engineer (\$5,016,723) is higher than the average lifetime earnings of an NFL Player(\$3,010,000) and an average MLB player (\$2,912,000). Furthermore, for the elite engineers have higher average career earnings than NBA players \$13,533,236 vs. \$12,027,456. New Orleans Saints cheerleaders are forbidden from eating in the same restaurant as any NFL player and if they are already dining at a restaurant and an NFL player comes in after, the cheerleaders are required to leave. the NFL uses tracking chips embedded in players shoulder pads to record a players every move on the field. They are also located inside footballs to track: velocity, rotation ,ball speed and location. The National Football League (NFL) is a professional American football league consisting of 32 teams, divided equally between the National Football Conference (NFC) and the American Football Conference (AFC). The NFL is one of the four major professional sports leagues in North America, and the highest professional level of American football in the world. Following the conclusion of the regular season, six teams from each conference (four division winners and two wild card teams) advance to the playoffs, a single-elimination tournament culminating in the Super Bowl, which is usually held in the first Sunday in February, and is played between the champions of the NFC and AFC. The NFL was formed in 1920 as the American Professional Football Association (APFA) before renaming itself the National Football League for the 1922 season. The NFL agreed to merge with the American Football League (AFL) in 1966, and the first Super Bowl was held at the end of that season; the merger was completed in 1970.
Utterance 2	Isn't that crazy that it's such a short amount of time for something they drag out of 3 hours?	
Utterance 3	I know! They sure do milk it for all its worth.	
Utterance 4	ESPN won an emmy for creating the superimposed yellow first down line on tv. That has been a good addition to watching games. Emotion: Happy	
Emo-STSC	Yes it was a long time ago, I wonder how long it took. Emotion: Curious to dive deeper	
Emo-STSC-F	Yes i like tv too. Emotion: Happy	
Emo-ITDD	It is very true, but I think it's crazy that there is only 11 min of live gameplay! Emotion: Surprised	
ITDD	It's surprising that the average NFL game only has 11 min of live gameplay! Emotion: Surprised	
KnowledGPT	That is true, ESPN won an emmy for creating the superimposed yellow line. Emotion: Neutral	
STSC	I think that is a good point. Did you know that the NFL uses tracking chips embedded in players shoulder pads. Emotion: Curious to dive deeper	
CAiRE	yes i know that. Emotion: Neutral	
Emo-Transformer	Did you know that? Emotion: Curious to Dive deeper	
EmoKbGAN	Did you know that the average lifetime earnings of an engineer is higher than the average lifetime earnings of an NFL player Emotion: Curious to Dive deeper	
ChatGPT	Sure, it may seem crazy, but that's just the way it is. Interestingly enough, did you know that engineers actually earn more on average than NFL players? And that's not the only surprising fact about the NFL. Despite the short duration of live gameplay, the league is one of the most popular professional sports in North America and attracts a huge audience. So while it may be "milked" for all it's worth, it's also a source of happiness and excitement for many people. Emotion: Happy	

Table 8

Some sample responses generated by the baseline and proposed models on our test set from the CMU_DoG dataset. The sentence used from the corresponding knowledge base is highlighted in bold.

Utterance 1	i was really impressed with the cg and special effects	Knowledge Base: moviename iron man year 2008 director jon favreau genre superhero cast robert downey jr.as tony stark - iron man, an industrialist, genius inventor, and consummate playboy, he is ceo of stark industries and a chief weapons manufacturer for the u.s. military. terrence howard as james rhodey rhodes, a friend of stark's, and the liaison between stark industries and the united states air force in the department of acquisitions, specifically weapons development. jeff bridges as obadiah stane, stark's business second-in-command, mentor and friend who turns on him to overtake the company, eventually building a giant exosuit to fight stark. shaun toub as yinsen, stark's fellow captive, who grafts an electromagnet to stark's chest to keep the shrapnel shell shards that wounded him from reaching his heart and killing hi, and who helps stark build the first iron man suit. gwyneth paltrow as pepper potts, stark's personal assistant and budding love interest. rating rotten tomatoes: 94% and average: 7.7 by 10 metacritic score: 79 by 100 cinemascore: a iron man is a 2008 american superhero film based on the marvel comics character of the same name, produced by marvel studios and distributed by paramount pictures.1 it is the first film in the marvel cinematic universe(mcu). the film was directed by jon favreau, with a screenplay by the writing teams of mark fergus and hawk ostby and art marcum and matt holloway. it stars robert downey jr. as tony stark by iron man, alongside terrence howard, jeff bridges, shaun toub, and gwyneth paltrow. in iron man, tony stark, an industrialist and master engineer, builds a powered exoskeleton and becomes the technologically advanced superhero iron man
Utterance 2	yes , favreau did a good job. he didnt go for stylized comic frames	
Utterance 3	yes i agree. i can 't believe that rotten tomatoes gave a 7.7/10 i think it deserves a much higher score	
Utterance 4	yes me too right. it deserves a higher score. i love pepper potts acting Emotion: Happy	
Emo-STSC	yes, do you like superhero movies ? Emotion: Happy	
Emo-STSC-F	yes. i agree with you. Emotion: Neutral	
Emo-ITDD	what did you think about the ending? Emotion: Curious to Dive deeper	
ITDD	yes Emotion: Neutral	
KnowledGPT	pepper potts stark assistant Emotion: Neutral	
STSC	i agree Emotion: Neutral	
CAiRE	yes. i agree. Emotion: Neutral	
Emo-Transformer	i wonder i like tv. Emotion: Surprised	
EmoKbGAN	yes robert downey tony stark	
ChatGPT	I completely agree with you both! Iron Man was such an amazing movie, and the special effects were really impressive. Favreau did an excellent job bringing the character to life on the big screen. And I'm not surprised that you think it deserves a higher score - it's one of my all-time favorites too! Emotion: Happy	

Table 9

Some sample responses generated by the baseline and proposed models on our test set from Topical chat dataset. The sentence used from the corresponding knowledge base is highlighted in bold.

Utterance 1	the burn book was such a designed to fail method. but you knew it was gonna cause funny drama haha	Knowledge Base: when regina is finally made aware of cady's treachery, she retaliates by spreading the contents of her burn book all over the school, quickly inciting a massive brawl. to avoid suspicion, regina inserts a fake libel of herself in the book in order to blame cady, gretchen, and karen, the only female students not mentioned in the book. karen convinces principal duvall that they did not spread the book. duvall soon quells the fighting, and gathers all the girls in the school in the gymnasium. math teacher ms. norbury, whom the burn book defamed as a drug dealer, makes the girls face the ways they all treat each other, confess their transgressions, and apologize to each other and the teachers. when janis's turn comes, she defies norbury and confesses her plan to destroy regina with cady's help , and openly mocks regina with the support of the entire school. pursued by an apologetic cady, regina storms out and is struck by a school bus, breaking her spine. without any friends, shunned by aaron, grounded by her parents and despised by her peers at school, cady takes full blame for the burn book and becomes an outcast. after she makes amends with regina, cady's guilt soon dissolves and she returns to her old personality. as part of her punishment for lying and failing ms. norbury's class, she joins the mathletes in the state championship finals, and ends up winning the competition for her team after her opponent answers incorrectly. at the spring fling dance, regina's new boyfriend, shane oman, is elected king, while cady is elected queen. onstage, cady declares that all her classmates are wonderful in their own way, breaks her plastic tiara, and distributes the pieces to some other girls. she then reconciles with janis, damian, and aaron, and reaches a truce with the plastics.
Utterance 2	the drama the burn book caused was serious !. a math teacher got defamed as a drug dealer because of it.	
Utterance 3	regina getting hit by that bus was a really intenses scene	
Utterance 4	it was. can you believe cady took all the blame for the burn book ? Emotion: Disgusted	
Emo-STSC	i like joe pesci and daniel stern. Emotion: Happy	
Emo-STSC-F	yes. i did not know that. Emotion: Surprised	
Emo-ITDD	alright, sounds like a good one of my favorite movies. have you seen the movie ? Emotion: Curious to dive deeper	
ITDD	i love joe pesci and daniel stern. Emotion: Happy	
KnowledGPT	Yes, it is the burn book Emotion: Neutral	
STSC	it came out in 1990. Emotion: Neutral	
CAiRE	have you seen the movie ? Emotion: Curious to Dive deeper	
Emo-Transformer	i love movies. Emotion: Happy	
EmoKbGAN	Have a great evening!	
ChatGPT	I'm sorry, but I cannot generate a response with the requested emotion of "disgusted" as it goes against OpenAI's content policy of avoiding negative or harmful content. Please provide a different emotion for me to generate a response with. Emotion: Neutral	

Furthermore, our proposed model is capable of generating smooth and satisfactory sentences such as *"It was nice chatting with you"* or *"Have a great day"*. In the instance of the baseline models, the ITDD and KnowledGPT, we observe that while these models are capable of generating responses based on external knowledge, they lack the participation necessary for communication and as a result, the responses get a little monotonous. When compared to these systems, our developed STSC model gives a considerably good performance by generating highly relevant responses as shown in Tables 6 and 7 and Tables 8 and 9 on the Topical Chat and CMU_DoG dataset. From the examples given, it is also clear that the baseline CAiRE and Emo-Transformer approaches produce responses that are emotionally consistent but are not factually accurate since they lack the necessary knowledge and information. The dialogue becomes less interactive as a result. However, our proposed Emo-Transformer model outperforms the CAiRE model by generating more emotionally satisfying responses. There are also some cases where our model fails to generate desirable responses. Our manual analysis reveals that a large percentage of errors is shown by out-of-context ground truth responses.

In the example of Table 7, we may note that the ground truth response does not fall in line with the context utterances and hence, we observe a drop in the prediction of such responses. Therefore, we can conclude that there is still ample room for enhancing the models. To exhibit the efficacy of the dual decoder, we compare the outcomes of the first and second decoders. All the tables, viz. Tables 6–9 feature the response generated by the first decoder (Emo-STSC-F), demonstrating the improvement brought about by the second decoder. In both cases, the second decoder leverages more comprehensive knowledge than the first one. Similarly, for the document-grounded conversations dataset, we have shown examples for the above-discussed cases. Inconsistency in the generated responses can also be attributed to the fact that there are errors in the annotation of emotion labels for the utterances. As seen in Table 9 Utterance 4 in the example should be labelled with *'Surprised'* emotion which is here getting confused with the *'Disgusted'* emotion. We also show responses from ChatGPT in Tables 6–9. We use the following instruction for this EKGDG task: Would you like to attempt generating a dialogue turn? You can produce only one turn in response to mine, and please make sure to incorporate the knowledge of XXX

for reference and convey a particular emotion *YYY* in your generated response. Although the responses generated by ChatGPT were deemed to be of high quality, they did not match the gold responses exactly. The model tended to generate additional information, which included content from its parametrized knowledge injected during pretraining.

5.5. Ablation study

To evaluate how the individual modules influence the efficiency of the model (Tables 4 and 5), we compare our proposed approach with the following variants:

(i) **STSC-coattnB**: We conduct experiments with only the Self-Attention Block for encoding the dialogue and knowledge sentences. For this experiment, we simply combine the utterance and knowledge encoding rather than employing the Step-wise Co-attention Block. We decode the responses using the dual decoder as proposed in Section 3.4. This is done to show how effective Step-Wise Co-attention Block is in encoding all of the conversation's implicit references. We observe a significant decrease in both BLEU and F1 scores for the Topical Chat and CMU_DoG dataset.

(ii) **STSC-DD**: Dual decoding is essential for generating insightful responses while decoding the response. We remove this module, as explained in Section 3.4, to demonstrate its utility. Both BLEU and F1 scores are significantly lowered as a result of this.

(iii) **STSC-EC**: Our final ablation model is the emotion-independent setup which is similar to the proposed knowledge grounded STSC model as described previously. This model evaluates the performance of the Emo-STSC model without the emotion classifier. We see a significant drop in scores here as well for both the datasets.

5.6. Model comparison with state of the art

We give comparisons with some of the most recent state-of-the-art methodologies in relation to the open-domain dialogue system in Tables 4 and 5 to demonstrate the effectiveness of our proposed network. In order to conduct a thorough study, we contrast our suggested strategy with regard to all of the *viz.* tasks, EGDG, KGDG, and EKGDG.

For the first task of knowledge-grounded dialogue generation, we start by comparing our model with the state-of-the-art model ITDD [8]. The ITDD model has an F1-score of 0.23/0.19 and a BLEU score of 1.04/0.88. Our STSC model performs better than the ITDD model with a BLEU score of 1.14 on the test frequent dataset and an F1-score of 0.25/0.22 on the topical chat dataset, as shown in Table 4. On the CMU_DoG dataset (Table 5), ITDD obtains a BLEU score of 1.08 and an F1-score of 0.18. Our proposed framework, shows improvement in terms of F1-score. Later, we also compare our method with the pre-trained auto-regressive model, i.e., $GPT - 2_{finetune}$ compared to which our model shows improvement with respect to all metrics on both the datasets. The $GPT - 2_{finetune}$ model uses a straightforward method of concatenating knowledge information in their framework for generating the responses, whereas we use a step-wise co-attention based method that has a greater potential to gather correlations for better response generation. This is the main cause of the degradation in performance of the pretrained GPT-2 model.

As [36] proposes a double latent variable architecture, we use the readily accessible implementation³ to conduct experiments on our freshly constructed dataset. According to evaluation results, both of our suggested frameworks *STSC* and *Emo-STSC* perform better than the current methodology across the board. The decoder suggested in [36] is unable to produce specific and pertinent responses since it lacks proper knowledge infusion methods. Similar to this, our suggested co-attention based dual decoder model demonstrates an improvement of 5% in the F1 score on both datasets compared to [32].

When compared to the difference aware model, the quality of responses is improved by the capacity to produce responses with the effective knowledge representation from our suggested co-attention module. Finally, we compare our proposed approach with two other approaches [33,35] that fuses knowledge information by using the knowledge selection models and in weakly unsupervised settings for generating responses using the available implementation.⁴ The table demonstrates that on both the Topical Chat and CMU_DoG datasets, the step-wise fusion technique beats the two alternatives. The improvement is a result of the conversation's several layers of information extraction being successfully combined to promote precise and consistent responses. On the task of emotion-grounded dialogue generation, it is evident from Tables 4 and 5 that by using a binary classifier to fuse emotion information into transformer-based architectures, we see improvement in comparison to the ECM [37] and CAiRE [46] frameworks on both the datasets. This may be due to the fact that the emotion classifier, as opposed to the RNN-based ECM model, has stronger guidance and style transfer capabilities. Contrarily, CAiRE's performance somewhat deteriorates in comparison to our discriminator-based method. The plausible explanation is that the authors of CAiRE [46] exclusively depend on the information provided by the pretrained language model and do not use any feedback systems for the generation of emotional responses.

In terms of all metrics, our suggested framework, Emo-STSC, performs better than the individual models for the tasks of emotion and knowledge-based generation, especially because it can concentrate on the right emotion and facts because the information is sent directly to the decoder. Additionally, our suggested framework's step-wise co-attention module, which captures the hierarchy in dialogue, gives a better conversational context than the baseline models. By conducting this analysis, we can conclude that our proposed framework surpasses the existing methods that are based solely on knowledge or emotion by being able to generate responses that are centred on both knowledge and emotion. These responses can effectively interact and adjust with the mood and behaviour of the agent in an open-domain dialogue setting. Emo-STSC shows competitive performance on perplexity, BLEU, F1 and diversity scores as well as on fluency, adequacy, emotional accuracy, and knowledge relevance. This suggests that a co-attention mechanism on the encoder side does help improve the generation process with a dual decoder and a classifier network to guide the training process according to relevant emotions. On our Topical Chat and CMU_DoG datasets, we see improvements of almost 2%–3% in terms of F1 scores from the existing knowledge and emotion-based approaches as designed and described in Section 4.2.

Additionally, the original paper [10] proposed a transformer-based model. It should be noted, however, that they, like us, did not focus on modelling the hierarchy between the context utterance and instead simply concatenated the context utterances and passed them as a single sequence into the transformer model. They discovered an F1-score of 0.22/0.20 on the frequent and rare test sets, respectively. On our proposed (Emo-STSC) model, we achieve a score of 0.25/0.22 for our task of emotion and knowledge controlled dialogue generation.

6. Conclusion and future work

In this paper, we have proposed an end-to-end neural network architecture for modelling knowledge-grounded conversations based on the dynamic co-attention network and an emotion classifier to model the emotions associated with each utterance. Our STSC model consists of a co-attention encoder which learns co-dependent representations of the utterances and of the relevant knowledge document, and a two-pass decoder which iteratively estimates the response.

⁴ <https://github.com/neukg/KAT-TSLF>; <https://github.com/zhaoxlpu/KnowledGPT>

³ <https://github.com/nlpucan/ZRKGK>

We have empirically shown that the iterative quality of the model helps in readdressing the predictions by picking up from the initial local maximum. We evaluate our models on two datasets, Topical Chat dataset [10] and Document Grounded Conversations Dataset [9]. Both automatic and human-based metrics show encouraging performance when generating emotionally relevant and factually correct responses on both datasets.

While transformer architectures like BERT have achieved remarkable success in various NLP tasks, especially classification, adapting them for generation tasks can be less straightforward. BERT, by design, is more attuned to classification tasks, and utilizing its pretrained weights directly for generation might not always yield optimal results. Therefore, while the integration of transformer-based embeddings holds potential, we approach this with caution, especially considering the potential inefficiencies of directly leveraging pretrained BERT weights for our generation-focused task. In future iterations, we aim to explore strategies to adapt transformer-based models more effectively for response generation, possibly by augmenting them with other architectures or through specialized fine-tuning.

Many natural language processing tasks require commonsense knowledge [63]. Because commonsense knowledge is the collection of background information people are meant to know and utilize during a discussion in open-domain conversational systems. In the future, we plan to leverage affective commonsense knowledge graphs, such as SenticNet [64], to investigate whether these can better comprehend the background information of a given utterance and make the response generation easier. In our experimental data usually, there are two agents involved between whom the conversation takes place. Capturing their dynamic personality in a dialogue introduces more variety in the responses. We aim to extend our work to datasets grounded on several other attributes like persona [65].

Finally, we plan to implement different reasoning strategies including parallel analogy, attempt to reduce the symbol grounding problem, explore different similarity measures, and endow the chatbot with intention awareness [66] in order to get one step closer to human-like dialogue systems.

CRedit authorship contribution statement

Deeksha Varshney: Resources, Methodology, Conceptualization. **Asif Ekbal:** Supervision, Methodology. **Erik Cambria:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

References

- [1] M. Firdaus, A. Ekbal, E. Cambria, Multitask learning for multilingual intent detection and slot filling in dialogue systems, *Inf. Fusion* 91 (2023) 299–315.
- [2] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, *Adv. Neural Inf. Process. Syst.* (2014) 3104–3112.
- [3] O. Vinyals, Q. Le, A neural conversational model, 2015, arXiv preprint arXiv:150605869.

- [4] I.V. Serban, A. Sordani, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3776–3783.
- [5] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, D. Yin, Knowledge diffusion for neural dialogue generation, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1489–1498.
- [6] M. Ghazvininejad, C. Brockett, M.W. Chang, B. Dolan, J. Gao, W.t. Yih, M. Galley, Knowledge-grounded neural conversation model, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018, pp. 5110–5117.
- [7] D. Varshney, A. Ekbal, M. Tiwari, G.P. Nagaraja, Emokbgan: Emotion controlled response generation using generative adversarial network for knowledge grounded conversation, *PLoS One* 18 (2) (2023) e0280458.
- [8] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, J. Zhou, Incremental transformer with deliberation decoder for document grounded conversations, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 12–21.
- [9] K. Zhou, S. Prabhume, A.W. Black, A dataset for document grounded conversations, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 708–713, <http://dx.doi.org/10.18653/v1/D18-1076>, URL <https://www.aclweb.org/anthology/D18-1076>.
- [10] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, Topical-chat: Towards knowledge-grounded open-domain conversations, in: *Proc. Interspeech* 2019, 2019, pp. 1891–1895, <http://dx.doi.org/10.21437/Interspeech.2019-3079>.
- [11] H. Xu, H. Peng, H. Xie, E. Cambria, L. Zhou, W. Zheng, End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization, *World Wide Web* 23 (2019) 1989–2002.
- [12] H. Golchha, M. Firdaus, A. Ekbal, P. Bhattacharyya, Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 851–860.
- [13] C.S. Wu, R. Socher, C. Xiong, Global-to-local memory pointer networks for task-oriented dialogue, 2019, arXiv preprint arXiv:190104713.
- [14] R.G. Reddy, D. Contractor, D. Raghu, S. Joshi, Multi-level memory for task oriented dialogs, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3744–3754.
- [15] A. Ritter, C. Cherry, W.B. Dolan, Data-driven response generation in social media, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2011, pp. 583–593.
- [16] L. Shang, Z. Lu, H. Li, Neural responding machine for short-text conversation, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1577–1586.
- [17] W.N. Zhang, L. Li, D. Cao, T. Liu, Exploring implicit feedback for open domain conversation generation, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018, pp. 547–554.
- [18] J. Ni, T. Young, V. Pandelea, F. Xue, V. Adiga, E. Cambria, Recent advances in deep learning based dialogue systems: A systematic survey, *Artif. Intell. Rev.* 56 (2023) 3055–3155.
- [19] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine, *Commun. ACM* 9 (1) (1966) 36–45.
- [20] D. Litman, S. Singh, M. Kearns, M. Walker, Njfun: a reinforcement learning spoken dialogue system, in: *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational Systems-Volume 3*, 2000, pp. 17–20.
- [21] J.D. Williams, S. Young, Partially observable markov decision processes for spoken dialog systems, *Comput. Speech Lang.* 21 (2) (2007) 393–422.
- [22] J. Schatzmann, K. Weilhammer, M. Stuttle, S. Young, A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies, *Knowl. Eng. Rev.* 21 (2) (2006) 97–126.
- [23] Z. Ji, Z. Lu, H. Li, An information retrieval approach to short text conversation, 2014, arXiv preprint arXiv:14086988.
- [24] R. Yan, Y. Song, H. Wu, Learning to respond with deep neural networks for retrieval-based human-computer conversation system, in: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 55–64.
- [25] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W.X. Zhao, D. Yu, H. Wu, Multi-turn response selection for chatbots with deep attention matching network, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1118–1127.
- [26] Q. Zhu, L. Cui, W. Zhang, F. Wei, T. Liu, Retrieval-enhanced adversarial training for neural response generation, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3763–3773.

- [27] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.Y. Nie, J. Gao, W.B. Dolan, A neural network approach to context-sensitive generation of conversational responses, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 196–205.
- [28] T.H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L.M. Rojas-Barahona, P.H. Su, S. Ultes, S. Young, A network-based end-to-end trainable task-oriented dialogue system, 2016, arXiv preprint [arXiv:160404562](https://arxiv.org/abs/160404562).
- [29] J. Du, W. Li, Y. He, R. Xu, L. Bing, X. Wang, Variational autoregressive decoder for neural response generation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3154–3163.
- [30] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston, Wizard of wikipedia: Knowledge-powered conversational agents, 2018, arXiv preprint [arXiv:181101241](https://arxiv.org/abs/181101241).
- [31] A. Madotto, C.S. Wu, P. Fung, Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1468–1478.
- [32] C. Zheng, Y. Cao, D. Jiang, M. Huang, Difference-aware knowledge selection for knowledge-grounded conversation generation, 2020, arXiv preprint [arXiv:200909378](https://arxiv.org/abs/200909378).
- [33] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, R. Yan, Knowledge-grounded dialogue generation with pre-trained language models, 2020, arXiv preprint [arXiv:201008824](https://arxiv.org/abs/201008824).
- [34] X. Zhao, W. Wu, C. Tao, C. Xu, D. Zhao, R. Yan, Low-resource knowledge-grounded dialogue generation, 2020, arXiv preprint [arXiv:200210348](https://arxiv.org/abs/200210348).
- [35] S. Liu, X. Zhao, B. Li, F. Ren, L. Zhang, S. Yin, A three-stage learning framework for low-resource knowledge-grounded dialogue generation, 2021, arXiv preprint [arXiv:210904096](https://arxiv.org/abs/210904096).
- [36] L. Li, C. Xu, W. Wu, Y. Zhao, X. Zhao, C. Tao, Zero-resource knowledge-grounded dialogue generation, *Adv. Neural Inf. Process. Syst.* 33 (2020) 8475–8485.
- [37] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: emotional conversation generation with internal and external memory, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, 2018, pp. 730–738.
- [38] X. Zhou, W.Y. Wang, Mojtalk: Generating emotional responses at scale, 2017, arXiv preprint [arXiv:171104090](https://arxiv.org/abs/171104090).
- [39] K. Wang, X. Wan, Sentigan: Generating sentimental texts via mixture adversarial networks, in: IJCAI, 2018, pp. 4446–4452.
- [40] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, E.P. Xing, Toward controlled generation of text, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1587–1596.
- [41] L. Logeswaran, H. Lee, S. Bengio, Content preserving text generation with attribute controls, *Adv. Neural Inf. Process. Syst.* (2018) 5103–5113.
- [42] Z. Song, X. Zheng, L. Liu, M. Xu, X.J. Huang, Generating responses with a specific emotion in dialog, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3685–3695.
- [43] S. Ghosh, M. Chollet, E. Laksana, L.P. Morency, S. Scherer, Affect-lm: A neural language model for customizable affective text generation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 634–642.
- [44] N. Asghar, P. Poupard, J. Hoey, X. Jiang, L. Mou, Affective neural response generation, in: European Conference on Information Retrieval, Springer, 2018, pp. 154–166.
- [45] C. Huang, O.R. Zaiane, A. Trabelsi, N. Dziri, Automatic dialogue generation with expressed emotions, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 49–54.
- [46] Z. Lin, P. Xu, G.I. Winata, F.B. Siddique, Z. Liu, J. Shin, P. Fung, Caire: An end-to-end empathetic chatbot, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 13622–13623.
- [47] Q. Li, P. Li, Z. Ren, P. Ren, Z. Chen, Knowledge bridging for empathetic dialogue generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 10993–11001.
- [48] T. Niu, M. Bansal, Polite dialogue generation without parallel data, *Trans. Assoc. Comput. Linguist.* 6 (2018) 373–389.
- [49] Z. Yang, Z. Hu, C. Dyer, E.P. Xing, T. Berg-Kirkpatrick, Unsupervised text style transfer using language models as discriminators, *Adv. Neural Inf. Process. Syst.* (2018) 7287–7298.
- [50] Y. Ma, K.L. Nguyen, F.Z. Xing, E. Cambria, A survey on empathetic dialogue systems, *Inf. Fusion* 64 (2020) 50–70.
- [51] M. Amin, E. Cambria, B. Schuller, Will affective computing emerge from foundation models and general AI? A first evaluation on ChatGPT, *IEEE Intell. Syst.* 38 (2) (2023) 15–23.
- [52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Kaiser Ł, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* (2017) 5998–6008.
- [53] C. Xiong, V. Zhong, R. Socher, Dynamic coattention networks for question answering, 2016, arXiv preprint [arXiv:161101604](https://arxiv.org/abs/161101604).
- [54] M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: EMNLP, 2015.
- [55] I.V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, Y. Bengio, A hierarchical latent variable encoder–decoder model for generating dialogues, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3295–3301.
- [56] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *Openai Blog* 1 (8) (2019) 9.
- [57] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
- [58] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 67–72, URL <https://www.aclweb.org/anthology/P17-4012>.
- [59] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [60] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.
- [61] J.L. Fleiss, Measuring nominal scale agreement among many raters, *Psychol. Bull.* 76 (5) (1971) 378.
- [62] B.L. Welch, The generalization of students' problem when several different population variances are involved, *Biometrika* 34 (1/2) (1947) 28–35.
- [63] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Common sense computing: From the society of mind to digital intuition and beyond, in: Biometric ID Management and Multimodal Communication, 2009, pp. 52–259.
- [64] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of LREC, 2022, pp. 3829–3839.
- [65] L. Zhu, W. Li, R. Mao, V. Pandelea, E. Cambria, PAED: Zero-shot persona attribute extraction in dialogues, in: Proceedings of ACL, 2023, pp. 9771–9787.
- [66] E. Cambria, R. Mao, M. Chen, Z. Wang, S.B. Ho, Seven pillars for the future of artificial intelligence, *IEEE Intell. Syst.* 38 (6), pp. 62–69 (2023).