# A Cognitive-grounded Computational Model for Emotional Alignment During Conversations

**Anonymous ACL submission**

## Abstract

Aligning emotionally during a conversation means showing appropriate emotional reactions to what our interlocutors say and the emotions they share with us. These appropriate emotional reactions are dictated by social standards and ensure smooth and effective interactions. Based on a psychological framework, and adapting already existing models from cognitive modeling and NLP, we investigate the role in conversational dynamics of 1) social expectations over emotional reactions, 2) internal emotional state, and 3) dialog acts. We implement and compare graph-based models and deep learning models on the task of emotion prediction, employing categorical accuracy as the target metric and using MELD and DialyDialog as benchmarks. The results suggest that the internal emotional state and the dialog acts have an influence on the emotional reaction during conversations. These elements, however, did not show a significant impact within the deep learning model. Possible improvements to the models and insights on future directions are provided.

## 1 Introduction

Conversation is a sophisticated activity, essential in our daily lives, given the social nature of human beings. To have a successful conversation people unconsciously and gradually imitate interlocutors' way of speaking. This phenomenon, called alignment, happens on several levels, such as phonetic (Pardo, 2006), lexical (Brennan and Clark, 1996), and syntactic (Branigan et al., 2000).

Alignment is also present on the emotional level, to create a common communication ground and fully understand others' intentions on a deeper level, beyond the semantics of their speech. Emotional alignment is defined not only as the mirroring of others' emotions but also as the appropriate reaction to our interlocutors' expressed emotions (Damm et al., 2011). A framework describing the conversational dynamics involving emotions was created by the German psychologist Reinhard Fiehler in 2002. Fiehler defined emotions as "public phenomena in social situations" (Fiehler, 2002) whose usage follows precise rules: i) Emotion rules, which determine which emotion is expected to be experienced given a context; ii) Manifestation rules, which state that the expected emotion must be expressed by the interlocutor; iii) Correspondence rules, regulating the appropriate emotional reaction to another person's feelings. It is thus clear how important are the *expectations* between events and emotions, and between emotions and emotional reactions in our daily communication. Moreover, Fiehler defines a set of possible reactions interlocutors may have in front of unexpected, i.e. socially unacceptable, usage of emotions: from the *ignoring* strategy, where the speaker does not take into account the latest contribution of their interlocutor, to the *calling into question* strategy, where the speaker temporarily suspends the main conversation to react to the interlocutor's input actively. Emotions are therefore not only *elements of* communication, but their usage depends and has an impact on the communicative behavior.

For these reasons, our aim is to create a computational model representing Fiehler's framework and, given the relevance of expectations in communication, employ it in a *emotion prediction* task on dyadic conversations. Our investigation, however, is purely cognitive-modeling oriented. Given the impressive performance of the current NLP models for affective computing our goal is not to create a new approach or architecture to outperform the current state-of-the-art models, but rather to investigate the role of 1) social expectations, 2) interlocutors' internal emotional status, and 3) speech acts in influencing the interlocutors' emotional reaction.

In section 2 we will give an overview of the models used in NLP for emotion recognition and prediction, event representation and events-emotions

interplay. In section 3, we will describe the models we implemented and compared.

## 2 Related Work

In recent years much effort has been put into emotion *recognition* research, and little to no studies focused on emotion *prediction*. Most of the definable "emotion prediction" models were dedicated to conversational agents capable of appropriate empathetic responses (Colombo et al., 2019; Zandie and Mahoor, 2020). Predicting the emotional content of the upcoming utterance based on the contextual information only is not an easy task. However, studies like Wen et al. (2021) accepted the challenge: Wen and colleagues proposed the implementation of interlocutors' personality representation combined with the emotions shared in the previous conversational turn and the general topic of the conversation. While the emotions are represented as 3-dimensional vectors representing Valence, Arousal, and Dominance (VAD space), the personalities are represented as 5-dimensional vectors having as features the strength in Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism of the talking character. These vectors are then used to predict the emotional response of each speaker, combining the emotion at the previous conversational turn, the personality of the speaker, and the emotion transition variation based on the utterance semantics. However, Wen and colleagues' model does not consider the interlocutor's emotional response which, as clearly stated in Fiehler (2002), has a big part in influencing the communication dynamic.

In Liang et al. (2022), a broader approach was adopted, utilizing a multimodal model to choose the suitable emotional response and generate relevant responses. The model comprises two main sections. The first part is a graph-based encoder that encompasses vector representations for the speaker, transcriptions of the utterances, emotions, facial expressions, and audio of the utterances. This information is inputted into a Feedforward Neural Network, which produces a prediction for the forthcoming emotion. The second component, called the "Emotion-Personality-Aware Decoder," utilizes the graph-enhanced representation, the predicted emotion, and the personality of the current speaker as inputs to generate a response that is coherent in terms of both content and emotion. The model was evaluated on MELD (Poria et al., 2019), and it outperformed state-of-the-art algorithms.

Even if our study is configured as an emotion prediction task, studies on emotion recognition in conversations are extremely valuable for finding different approaches and resources.

For example, Majumder et al. (2019) presented an architecture relying on a sequence of recurrent neural networks with gated recurrent units (GRUs). The model captures four key aspects of the conversation: the context of each utterance, each speaker's emotional state, the relevant context for the current utterance, and the listener's internal state based on the speaker's utterance. The model was evaluated on IEMOCAP (Busso et al., 2008) and AVEC (Schuller et al., 2012), and compared against baselines using LSTMs and GRUs. Results showed that the proposed model achieved the highest F1 scores on IEMOCAP and the lowest MAE on AVEC. These findings highlight the importance of representing each speaker's internal state and the impact of interactions for successful emotion recognition. In Zhu et al. (2021) a representation of the conversational context is used within a system composed of a language model, a transformer, and general knowledge retrieved from ATOMIC (Sap et al., 2019), a knowledge graph where events are linked to each other by specific sets of relationships (causes, effects, agents, themes, etc.). ATOMIC is also used by Sabour et al. (2022), who proposed a model to generate empathetic responses that combines a module to recognize the interlocutor's emotion and take track of its fluctuations during the interaction, and a component to retrieve the knowledge regarding the conversational context.

For developing and testing models for emotion recognition and prediction in conversational scenarios, various datasets have been created, such as DailyDialog (Li et al., 2017), MELD (Poria et al., 2018), Empathetic Dialogues (Rashkin et al., 2018b), Story commonsense (Rashkin et al., 2018a). These datasets differ in their size, number of emotion labels, and nature (written chat messages or transcriptions of oral conversations). DailyDialog also presents the annotation for *dialog acts*, representing the communication functions of the utterances.

Dialog acts are the fundamental units of meaning used by interlocutors to deliver their intentions and convey information. Dialog acts are a broad category of communicative actions such as posing queries, giving clarifications, asserting claims, issuing instructions, voicing opinions, requesting

things, and so on. The relationship between emotions and dialog acts is investigated in Cao et al. (2021). Analyzing DailyDialog and dyadic MELD materials, the authors found a series of causal relationships between dialog acts and emotions, for example happiness and thanking, fear and interruption, sadness and apology, surprise and asking questions, disgust and statement-opinion and wh-questions, anger and action-directive.

A similar investigation was conducted on Japanese by Ihasz and Kryssanov (2018), who analyzed the co-occurrences of emotions and dialog acts in the transcriptions of conversations during online gaming sessions, quantified using the normalized pointwise mutual information (*npmi*). As in English, consistent and predictable dialog act-emotion pairs were found.

On the other hand, cognitive modeling research presents useful insights regarding expectations and events. In Chersoni et al. (2019) the authors proposed a model based on the notion of Generalized Event Knowledge (Metusalem et al., 2012), which theorizes that people have mental representations of events and their typical participants, creating thus networks of expectations. Chersoni and colleagues created a graph representing this knowledge and used such information to create lists of predictions over upcoming elements in sentences.

The interplay between emotions and events was the focus of Lee et al. (2013) where emotions were found to be pivot events between causes and effects. Emotions can be thus seen, and therefore treated as, events themselves.

Putting together the different theories and previous models, the aim of this study is to create a cognitive-grounded graph-based model to investigate 1) the role of social expectations in driving emotional reactions in conversations, 2) if and how the internal emotional status influences the emotional reactions, and 3) whether expectations over the dialog act emotional reactions are delivered with have an impact on the emotional reactions. To extend the investigation, we use the same components within a deep learning architecture too.

# 3 Method

## 3.1 Datasets

- **MELD** (Poria et al., 2019): a multimodal dataset with dialogues selected from the TV series "Friends". For each utterance, transcription of the utterance, a video clip, emotion annotation (neutral, joy, fear, surprise, sadness, anger, and disgust.), and sentiment annotation (positive, negative, and neutral) are provided.
- **DailyDialog** (Li et al., 2017): large dataset of written dyadic conversations. Each utterance text is annotated for emotion (neutral state, anger, happiness, fear, disgust, surprise, and sadness), dialog act (inform, question, directive, and commissive), and topic (work, attitude and emotion, health, culture and education, tourism, politics, finance, relationship, school life, and ordinary life.).

For both datasets, the original split into training, validation, and test set was followed to train the models, asses the hyperparameters, and compare the models' performances, respectively.

## 3.2 Data pre-processing

We focus on the dyadic conversations subset of MELD and merge together those consecutive turns produced by the same speaker. The utterances were concatenated one after the other and vectorized as a whole using BERT Large pre-trained vectors (Devlin et al., 2018), specifically the embedding corresponding to the special token `[CLS]`. The emotions of these turns were approximated as the most frequent one of the sequence or, if multiple emotions appeared with the same frequency, the one expressed in the last utterance was used as representative of the whole sequence. Also, since DailyDialog conversations are annotated for dialog act too, a simple feedforward neural network was implemented and trained to classify an utterance as *information, question, directive, commissive* and extend such annotation to MELD utterances.

Only conversations with at least 3 turns were selected from MELD and DailyDialog. The resulting datasets were however heavily imbalanced with the neutral emotion being up to 50% of the labels in MELD and 80% in DailyDialog. The datasets were then further modified, excluding those conversations where the neutral emotion represented 70% or more of the turns. The size of the resulting datasets was consequently heavily reduced.

## 3.3 Models

In the present study, we propose and compare 4 main models:

1. **GEmK** (baseline): a graph-based model to represent the social expectations-driven emotional reactions;

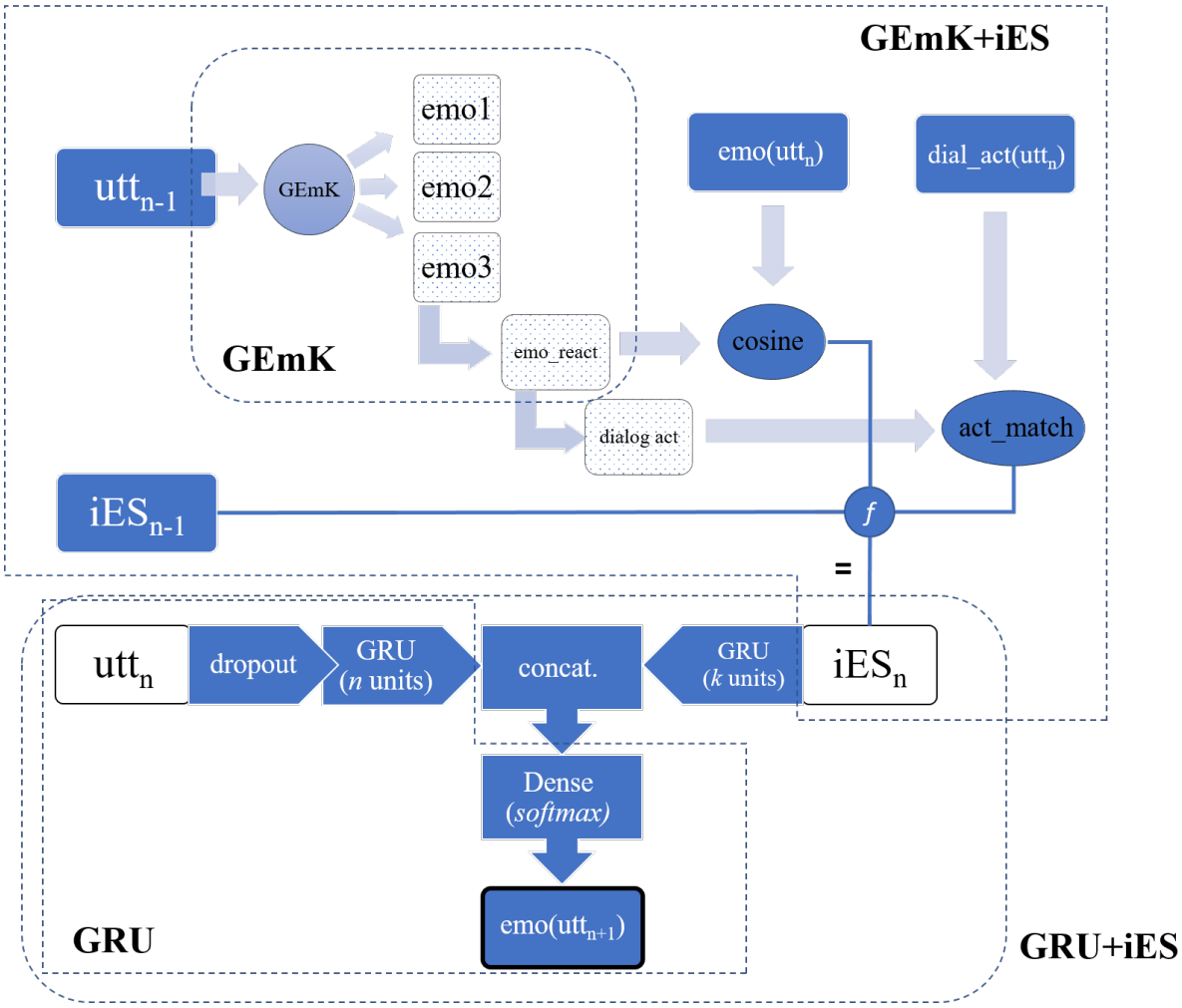


Figure 1: Architecture employed in GRU+iES.

2. **GEmK + iES**: a graph-based model to represent the emotional reactions emerging from the interaction between social expectations and internal emotional status;

3. **GRU**: a deep learning model (Gated recurrent units neural network, as in Majumder et al. (2019)) to represent the utterance-driven emotional reactions;

4. **GRU + iES**: a deep learning model to represent the emotional reactions emerging from the interaction between utterances and internal emotional status.

Models GEmK and GEmK + iES are based on a graph called *Generalized Emotional Knowledge*, which represents the social expectations between events and emotions and between emotions and emotional reactions, having thus events and emotions as nodes, *cause*, *effect*, and *emotional reaction* as edge labels and using the Local Mutual Information (LMI) between elements as edge values.

The graph was built as follows: from ATOMIC, events containing emotion-related terms (e.g., *PersonX feels happy*, *PersonX is disgusted*) were collected and used as pivot-events. The events these emotional events were connected to as their causes or effects (i.e., relations *xNeed, xEffect, xWant, xReact*) were converted to 1024-dimensional vectors using BERT Large. Similar nodes (namely, vectors with a cosine similarity of 0.90 or higher) connected to the same emotions were merged, and the Local Mutual Information (LMI) between event and emotion was updated. The event-emotion connections were then enriched with information extracted from Empathetic Dialogues.

The emotions that the different events were connected to were included in the graph as 3-dimensional vectors, representing the valence, arousal, and dominance of such emotions, as in Wen et al. (2021) (see Table1). From MELD and

| Emotion | Vector |
|---|---|
| Neutral | [-0.01, -0.01, 0.0] |
| Happiness | [0.81, 0.51, 0.46] |
| Sadness | [-0.63, -0.27, -0.33] |
| Anger | [-0.51, 0.59, 0.25] |
| Fear | [-0.62, 0.82, -0.43] |
| Disgust | [-0.60, 0.35, 0.11] |
| Surprise | [0.40, 0.67, -0.13] |

Table 1: Representation of emotions in the VAD space, following the implementation of Wen et al. (2021), except for Neutral, whose vector has been modified from the original.

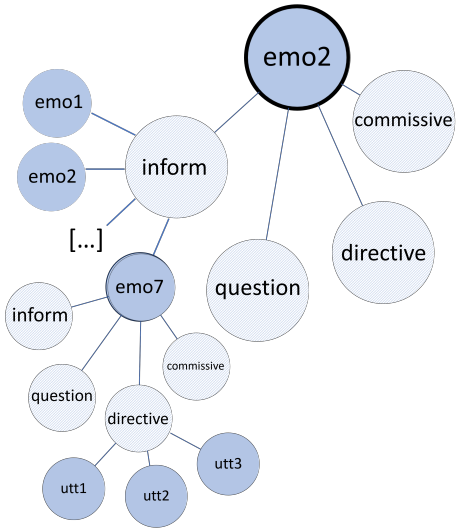


Figure 2: Focus on dialog acts structure within GEmK.

DailyDialog information regarding emotional reactions to expressed emotions was extracted: emotions found in two consecutive turns were linked together, with the second being the *emotional reaction* of the previous one. Recurrences of a pair of emotions were used to increase the LMI between the (nodes representing the) two emotions.

As mentioned in Section 2, corpora studies found that specific dialog acts tend to co-occur with specific emotions. This interplay is represented in GEmK through a sub-graph where dialog acts of statements and dialog acts of replies are mediated by the expectations between emotions (Figure 2). More specifically, each emotional state in GEmK is connected to the four action units annotated in DailyDialog. Each action unit is in turn connected to seven nodes representing the seven emotions the present models focus on. This layer of connection indicates the emotional reaction to the emotion found at the upper level and delivered through that specific action unit. Each emotional state is then connected to four nodes representing the action units, namely the way the emotional reaction is delivered. The edges are computed as MI taking into account the frequencies of the first emotion-dialog act set of pairs, the frequencies of emotion-emotional reaction, and the frequencies of the second layer of emotion-dialog act pairs.

As mentioned before, given the communicative importance of the interlocutors' expectations between expressed events or emotions and emotional reactions, we formulate our investigation as an emotion prediction task. To predict the upcoming emotional reaction each model worked as follows:

In **GEmK**, once an utterance is found, it is converted into a vector and compared with the nodes present in the graph to find the most similar one using cosine similarity. The emotion with the strongest connection to that node is used as the pivot element to predict the emotion that will be expressed by the interlocutor. This selection is based on the LMI between the two nodes connected through the *emotional reaction* labeled edge.

In **GEmK + iES** the prediction is computed the same way as GEmK but with the influence of the internal emotional state of the interlocutor who is about to talk. The internal emotional state (iES) is modeled as a 3-dimensional vector, like the emotions in GEmK, and initialized with the emotion of the first utterance of each interlocutor. At each turn, it is updated as follows:

$$iES_{(t)} = iES_{(t-1)} * (cos(emo\_react_{(t-1)}, emo_{(t)}) + act\_match) \quad (1)$$

Where $iES_{(t-1)}$ is the emotional state of that interlocutor at the previous turn. $cos(emo\_react_{(t-1)}, emo_{(t)})$ models how close the expectations between the emotional reaction found appropriate at the previous turn and the actual emotional reaction are, expressed as the cosine similarity between the two emotion vectors. Finally, $act\_match$ is the output of a function that returns 0 if the dialog act of the current utterance matches the expectations at the previous turn, 0.5 if they belong to the same macro-group (i.e., exchange of information or accepting/rejecting offers), 1 if the expected act and the current one completely mismatch. The dialog act expectations were computed at the previous conversational turn, following the sub-graph of expected emotions and related dialog acts.

Then, at each turn, a prediction on the emotional reaction by the interlocutor is computed:

$$Pred\_react = iES_{(t-1)} + emo\_react_{(t)} \quad (2)$$

where $iES_{(t-1)}$ is the next speaker's emotional state at the previous turn, and $emo\_react_{(t)}$ is the socially appropriate emotional reaction given what is being said at the current turn.

In order to investigate which elements are valuable in predicting the upcoming emotion, we perform an ablation study including different combinations of the components present in (1) and (2) taking track of the performance variations. Finally, the model with the highest categorical accuracy is used to produce the iES vectors to be used in GRU + iES.

**GRU** is a recurrent neural network with gated recurrent units, implemented using *Keras*[1]. The basic version is trained to predict the emotion of the upcoming turn given a vector representing the current utterance. The network presents an $Input$ layer, a $dropout$ layer, a $GRU$ layer, and finally a $Dense$ layer with $softmax$ as the activation function. The network has $Adam$ as an optimizer and the monitor metric for the training is the $categorical\ accuracy$.

**GRU + iES** presents an architecture similar to GRU, but enriched with two more input layers: one having the iES computed at each turn from GEmk + iES and one having the act of the current utterance, encoded as a 4-d one-hot vector. The three $Input$ layers are connected to three independent $GRUs$, whose outputs are then concatenated and fed into the $Dense$ layer. Only the $Input$ layer handling the utterance vector is followed by a $dropout$ layer, to mitigate the dimensionality difference among the three types of input.

For all the models using GRUs, the hyperparameters such as number of units, dropout value, and learning rate were tuned through a Bayesian Optimization algorithm implemented using *hyperopt*[2].

## 4 Results

Table 2 shows the performances of the different models on MELD and DailyDialog (now on indicated as DD) test sets in terms of categorical accuracy.

[1]https://keras.io/

[2]http://hyperopt.github.io/hyperopt/

| Model | Dataset | Accuracy |
|---|---|---|
| **GEmK** (baseline) | MELD | 0.24 |
| | DD | 0.31 |
| **GEmK+iES** | MELD | 0.31 |
| | DD | 0.41 |
| **GRU** | MELD | 0.24 |
| | DD | 0.49 |
| **GRU+iES** | MELD | 0.33 |
| | DD | 0.55 |

Table 2: Categorical accuracy of the 4 models in comparison.

| Model | Dataset | Accuracy |
|---|---|---|
| **GEmK** (baseline) | MELD | 0.26 |
| | DD | 0.49 |
| **GEmK+iES** | MELD | 0.42 |
| | DD | 0.80 |
| **GRU** | MELD | 0.40 |
| | DD | 0.81 |
| **GRU+iES** | MELD | 0.42 |
| | DD | 0.81 |

Table 3: Categorical accuracy of the 4 models on the original, imbalanced version of MELD and DailyDialog test sets.

**Basic models vs. iES**. As shown in Table 2, we have an improvement in the accuracy levels when including the iES component within the GEmK framework across both datasets. The target metric increases from $0.24$ to $0.31$ on MELD and from $0.31$ to $0.41$ on DailyDialog. Similarly, the GRU model sees an improvement on MELD, from $0.24$ to $0.33$, and from $0.49$ to $0.55$ on DailyDialog.

As reported in Table 4 (and Table 6 in Appendix A), the presence of iES does not modify the models' general prediction patterns, but it helps in improving the prediction performances. For example, in GEmK the easiest labels to predict are neutral and happiness, while the most difficult are sadness and surprise. GEmK+iES reports the same trend but with higher accuracy scores, and with anger as the new second-best predicted label.

**GEmK vs. GRU**. Averaging over the models' performances on MELD and DailyDialog, GRU models have higher categorical accuracy than GEmK models, namely $27.5\%$ vs $36.5\%$ for the basic models and $36\%$ vs $44\%$ for the iES models. However, on MELD the two models show very similar performance, with GEmK and GRU showing the same accuracy and GEmK+iES and GRU+iES

| | GEmK | | | GEmK+iES | | |
|---|---|---|---|---|---|---|
| | precision | recall | f1-score | precision | recall | f1-score |
| Neutral | 0.32 | 0.51 | 0.39 | 0.34 | 0.59 | 0.43 |
| Anger | 0.19 | 0.08 | 0.12 | 0.36 | 0.25 | 0.30 |
| Disgust | 0.12 | 0.08 | 0.10 | 0.12 | 0.08 | 0.10 |
| Fear | 0.04 | 0.12 | 0.06 | 0.25 | 0.12 | 0.17 |
| Happiness | 0.21 | 0.26 | 0.23 | 0.32 | 0.26 | 0.29 |
| Sadness | 0.17 | 0.05 | 0.08 | 0.20 | 0.05 | 0.08 |
| Surprise | 1.00 | 0.03 | 0.06 | 0.17 | 0.10 | 0.12 |
| | **GRU** | | | **GRU+iES** | | |
| | precision | recall | f1-score | precision | recall | f1-score |
| Neutral | 0.20 | 0.03 | 0.06 | 0.00 | 0.00 | 0.00 |
| Anger | 0.50 | 0.08 | 0.14 | 0.30 | 0.29 | 0.30 |
| Disgust | 0.00 | 0.00 | 0.00 | 1.00 | 0.08 | 0.14 |
| Fear | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Happiness | 0.20 | 0.66 | 0.30 | 0.35 | 0.74 | 0.47 |
| Sadness | 0.35 | 0.22 | 0.27 | 0.24 | 0.12 | 0.16 |
| Surprise | 0.29 | 0.27 | 0.28 | 0.31 | 0.21 | 0.25 |

Table 4: GEmK, GEmK+iES, GRU, and GRU+iES performances in terms of precision, recall and f1-score for each emotional category on MELD.

differing only for a $2\%$.
Focusing on the model performances on different emotional categories, it is possible to notice that, even if GEmK models show a generally lower precision, recall, and f1-score, they have a more uniform prediction distribution: GRU models have good performances on anger, happiness, and surprise, but have low performances (reaching zero) on neutral, disgust, and fear, registering a gap of $47\%$ in f1-score between happiness and fear. On the other hand, the strongest performance difference with GEmK models is a $35\%$ gap between neutral and sadness.

### 4.1 Ablation study on GEmK + iES

As mentioned before, GEmK+iES was subject to an ablation study, where different components were dynamically included and combined. On MELD, the best accuracy $(0.31)$ was reached by 4 different versions, all of them having three elements in common:

- the inclusion of the previous internal emotional status ($iES_{(t-1)}$)
- the inclusion of the (mis)match evaluation on the dialog act ($act\_match$)
- the exclusion of the socially-driven emotional reaction

Similarly, on DailyDialog the best model was the one employing the internal emotional status and the dialog act matching function only.

### 4.2 Comparison with SOTA models

Although our goal is not to propose a competitive NLP model for emotion prediction, but rather implement a cognitive-inspired model to analyze the contribution of psychological and linguistic elements on emotional reaction, a comparison of GEmK+iES and GRU+iES with previous ones is still needed. We focused on:

- **Emo-HRED**, proposed by Lubis et al. (2018) but following the implementation in Liang et al. (2022);
- **HGNN+BART** by Liang et al. (2022);
- **PET-CLS** by Wen et al. (2021);
- **RoBERTa** proposed by Liu et al. (2019) but following the implementation in Wen et al. (2021).

To ensure an informative comparison we report the performances on the imbalanced dataset in terms of weighted average F1 (Table 5). Both GEmK+iES and GRU+iES are outperformed by Emo-HRED, HGNN+BART, and PET-CLS, with the first two outperforming the proposed models on both MELD

| Model | W-avg F1 | |
|---|---|---|
| | **MELD** | **DD** |
| Emo-HRED | 31.03 | 79.67 |
| HGNN+BART | 36.73 | 83.88 |
| PET-CLS | 42.4 | - |
| RoBERTa | 28.7 | - |
| GEmK+iES | 30 | 74 |
| GRU+iES | 29 | 75 |

Table 5: Comparison with state-of-the-art models, with results as reported in the original papers. The performances are evaluated on the original version of MELD and DailyDialog.

and DailyDialog. The models described in this paper still however achieve performances comparable with the ones presented in Liang et al. (2022), and slightly outperform RoBERTa as reported by Wen et al. (2021).

To sum up the main findings:

1. Our graph-based baseline model, GEmK, achieved the same performance as the basic GRU model on MELD, and it was outperformed by the deep learning architecture on DailyDialog;

2. The internal emotional status (iES) improved the performances of both GEmK and GRU on both corpora. However, GEmK received the overall strongest benefit from the employment of iES;

3. The ablation study performed on GEmK+iES showed that the inclusion of the match evaluation function on dialog acts and internal emotional status of the previous turn are the elements leading to the best-performing models. The socially-driven emotional reaction and the match evaluation function on expected emotional reactions were inconsistently part of the computation of the best GEmK+iES model on MELD and consistently excluded on DailyDialog.

## 5 Discussion

As mentioned in Section 2, in the past years many NLP models employed a representation of the emotional state of the interlocutors to improve their performance in emotion recognition and prediction. Our study confirms the relevance of such a representation. In fact, the usage of iES improves both the graph-based and the deep learning model, and the inclusion of the emotional status of the previous turn (psychologically defined as the "emotional inertia") in the calculation of the current one, shows how the sole semantic information is not sufficient to predict emotional reactions, and a psychological-oriented representation is needed.

Unexpectedly, the two elements explicitly aiming at representing Fiehler's framework, i.e., the influence of social expectations in leading the emotional reaction and the influence of (un)met expectations on interlocutors, turned out to be not beneficial for the models. However, as explained in Section 3.3, the information about the expectations on dialog acts was emotion-driven. In GEmK we find connections of the type $emotion - dialog\_act \rightarrow emotion - dialog\_act$, so, although indirectly, whether the expectations over the emotional reaction are met is represented in the *act_match* function, and it thus still influences the internal emotional state computation, and hence the prediction of the emotional reaction.

## 6 Conclusion

In the present paper, we investigated the role of different cognitive-based elements in modeling conversational dynamics. The results suggested that having a representation of the interlocutors' emotional state and their expectations over the dialog acts used by their conversational partners helps in predicting the emotion of the upcoming utterance. These elements' contribution, however, is not significant in deep learning models, where the performance did not benefit from their usage as much as the graph-based model.

## 7 Limitations

GEmK models: even if the emotional knowledge graph was built upon one knowledge graph material and three conversational datasets, it would have benefited from a greater amount of information. Also, in future work it may be appropriate to modify the formulas for emotion prediction in GEmK + iES (1 and 2), and make them fully parametric, to express a more granular contribution of each element.

GRU models: although more technically advanced than GEmK models, the deep-learning models here proposed are still quite basic in their architecture, and a possible improvement would be the inclusion of an attention mechanism, that proved to help in

the emotion recognition and prediction tasks.

## References


Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Shuyi Cao, Lizhen Qu, and Leimin Tian. 2021. Causal relationships between emotions and dialog acts. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Emmanuele Chersoni, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and C-R Huang. 2019. A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, 25(4):483–502.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

Oliver Damm, Karoline Malchus, Frank Hegel, Petra Jaecks, Prisca Stenneken, Britta Wrede, and Martina Hielscher-Fastabend. 2011. A computational model of emotional alignment. In *5th Workshop on Emotion and Computing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Reinhard Fiehler. 2002. How to do emotions with words: Emotionality in conversations. *The verbal communication of emotions*, pages 79–106.

Peter Lajos Ihasz and Victor Kryssanov. 2018. Emotions and intentions mediated with dialogue acts. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pages 125–130. IEEE.

Sophia Yat Mei Lee, Huarui Zhang, and Chu-Ren Huang. 2013. An event-based emotion corpus. In *Workshop on Chinese Lexical Semantics*, pages 635–644. Springer.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022. Emotional conversation generation with heterogeneous graph neural network. 308:103714.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. 33(1):6818–6825.

Ross Metusalem, Marta Kutas, Thomas P Urbach, Mary Hare, Ken McRae, and Jeffrey L Elman. 2012. Generalized event knowledge activation during online sentence comprehension. *Journal of memory and language*, 66(4):545–567.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Preprint*, arxiv:1810.02508 [cs].

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018b. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019.

Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456.

Zhiyuan Wen, Cao Jiannong, Yang Ruosong, Liu Shuaiqi, and Shen Jiaxing. 2021. Automatically select emotion for response via personality-affected emotion transition. *arXiv preprint arXiv:2106.15846*.

Rohola Zandie and Mohammad H Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

# A Appendix

| | **GEmK** | | | **GEmK+iES** | | |
|---|---|---|---|---|---|---|
| | precision | recall | f1-score | precision | recall | f1-score |
| Neutral | 0.32 | 0.59 | 0.41 | 0.40 | 0.50 | 0.44 |
| Anger | 0.04 | 0.05 | 0.04 | 0.16 | 0.19 | 0.17 |
| Disgust | 0.05 | 0.11 | 0.07 | 0.16 | 0.42 | 0.23 |
| Fear | 0.04 | 0.40 | 0.08 | 0.05 | 0.40 | 0.08 |
| Happiness | 0.52 | 0.22 | 0.31 | 0.62 | 0.44 | 0.51 |
| Sadness | 0.15 | 0.05 | 0.07 | 0.19 | 0.15 | 0.16 |
| Surprise | 1.00 | 0.04 | 0.07 | 0.07 | 0.04 | 0.05 |
| | **GRU** | | | **GRU+iES** | | |
| | precision | recall | f1-score | precision | recall | f1-score |
| Neutral | 0.31 | 0.30 | 0.30 | 0.49 | 0.33 | 0.40 |
| Anger | 0.70 | 0.55 | 0.62 | 0.61 | 0.72 | 0.66 |
| Disgust | 0.42 | 0.25 | 0.31 | 0.71 | 0.31 | 0.43 |
| Fear | 0.50 | 0.36 | 0.42 | 0.47 | 0.50 | 0.48 |
| Happiness | 0.46 | 0.37 | 0.41 | 0.60 | 0.43 | 0.50 |
| Sadness | 0.46 | 0.64 | 0.53 | 0.56 | 0.53 | 0.55 |
| Surprise | 0.50 | 0.61 | 0.55 | 0.48 | 0.69 | 0.57 |

Table 6: GEmK, GEmK+iES, GRU, and GRU+iES performances in terms of precision, recall and f1-score for each emotional category on DailyDialog.