Identifying Distorted Thinking in Patient-Therapist Text Message Exchanges by Leveraging Dynamic Multi-Turn Context

Kevin Lybarger **Justin S. Tauscher** **Xiruo Ding**
Dror Ben-Zeev **Trevor Cohen**
{lybarger, jtausch, xiruod, dbenzeev, cohenta}@uw.edu
University of Washington, Seattle, WA

Abstract

There is growing evidence that mobile text message exchanges between patients and therapists can augment traditional cognitive behavioral therapy. The automatic characterization of patient thinking patterns in this asynchronous text communication may guide treatment and assist in therapist training. In this work, we automatically identify distorted thinking in text-based patient-therapist exchanges, investigating the role of conversation history (context) in distortion prediction. We identify six unique types of cognitive distortions and utilize BERT-based architectures to represent text messages within the context of the conversation. We propose two approaches for leveraging dynamic conversation context in model training. By representing the text messages within the context of the broader patient-therapist conversation, the models better emulate the therapist's task of recognizing distorted thoughts. This multi-turn classification approach also leverages the clustering of distorted thinking in the conversation timeline. We demonstrate that including conversation context, including the proposed dynamic context methods, improves distortion prediction performance. The proposed architectures and conversation encoding approaches achieve performance comparable to inter-rater agreement. The presence of any distorted thinking is identified with relatively high performance at 0.73 F1, significantly outperforming the best context-agnostic models (0.68 F1).

1 Introduction

Cognitive behavioral therapy (CBT) is an evidence based treatment applicable to a wide range of mental health conditions including depression, anxiety, addiction, bipolar disorder, and schizophrenia spectrum disorders (Yurica and DiTomasso, 2005; Hofmann et al., 2012). One primary clinical activity of CBT is the identification and re-framing of systematic errors in thinking, termed *cognitive distortions*, that create a skewed perception of reality (Beck, 1963). Cognitive distortions are known to exacerbate psychiatric symptoms without intervention (Dudley et al., 2016); however, there are many types of cognitive distortions (e.g., overgeneralization or catastrophizing), which can make identification and appropriate intervention by clinicians more complicated (Burns, 1980).

CBT has traditionally been administered through in-person office visits; however, there is increasing need for remote therapy options, to extend provider reach and increase access (Lin and Espay, 2021). Remote therapy options include internet-delivered therapy, application-based therapy, teletherapy, and text messaging (Lin and Espay, 2021; D'Arcey et al., 2020). There is growing evidence that asynchronous text-message-based exchanges between patients and therapists can augment conventional synchronous therapy and improve patient outcomes (D'Arcey et al., 2020). The expansion of text-message-based CBT provides an opportunity to develop clinician supports via novel natural language processing (NLP) methods that can guide patient treatment and assist in therapist training.

In this work, we explore the automatic identification and categorization of cognitive distortions in a corpus of text-message conversations between patients with serious mental illness and their therapists. Prior work identifying cognitive distortions in text treats each text sample (e.g. sentence or message) as an independent event without context. However, in this conversational paradigm, the preceding turns in the conversation may provide important contextual cues for recognizing distorted thinking. Here, we utilize state-of-the-art deep learning NLP methods to explore the role of conversation history in identifying cognitive distortions in patient-therapist text message exchanges. By identifying distorted thinking in text messages within the broader context of the dialogue, the dialogue-based prediction architectures emulate the real-world process of mental health clinicians who

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 126 - 136

account for conversation context when assessing for distortions. The dialogue-based architectures also mirror the cognitive distortion annotation process associated with the data set used in this work. We present multiple BERT-based architectures for identifying distortions in multi-turn conversations and propose methods for dynamically representing the conversation context. We demonstrate that leveraging the dialogue context and incorporating the proposed dynamic conversation context yields statistically significant performance improvement, reaching performance levels comparable to inter-rater agreement. Distorted thinking is identified in the text messages at 0.73 F1.

2 Related Work

There is a relatively small body of work exploring the automatic identification and categorization of cognitive distortions in user-generated text. Wiemer-Hastings et al. (2004) explored the identification of dysfunctional thoughts in 188 text examples from the cognitive distortion literature. The authors manually curated linguistic features that were used in a decision tree. Simms et al. (2017) annotated 459 Tumblr blogs for the presence of cognitive distortions. Features were extracted using the Linguistic Inquiry and Word Count (LIWC) tool (Tausczik and Pennebaker, 2010)[1], and several classifiers were explored with logistic regression (LR) achieving the best performance. Shickel et al. (2020) investigated the identification of cognitive distortions in online journal entries from college students and samples from crowdsourced participants prompted to give examples of defined distortion types. The authors investigated many classification architectures, including LR, Support Vector Machines (SVM), recurrent neural networks (RNN), convolutional neural networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The authors reported the highest performance using LR with with term frequency-inverse document frequency (TF-IDF) features. Shreevastava and Foltz (2021) explored the classification of 10 distinct cognitive distortions in 3,000 therapist question-answer samples. Several classifiers and feature encoding approaches were explored, and the best performance was achieved by an SVM operating on the SentenceBERT encoding, without fine-tuning BERT.

We explored the identification of cognitive distortions in patient-therapist text message exchanges and implemented the best performing models from Shickel et al. (2020) (LR with TF-IDF) and Shreevastava and Foltz (2021) (SVM with SentenceBERT without fine-tuning) as baselines. We found that fine-tuning BERT for multi-label classification achieves state of the art performance in our cognitive distortion prediction task, so we focus the experimentation in this work on BERT architectures. We are not aware of any cognitive distortion prediction work that leverages conversation history as context for identifying distorted thinking.

In this work, we identify cognitive distortions in text-based conversations, exploring the role of conversation history. This distortion prediction task shares similarities with other multi-turn conversational tasks, including retrieval-based dialogue response generation and question answering. Dialogue response and question answering are often approached using hierarchical architectures that first encode each turn, then aggregate the turn embeddings to create a conversation embedding, and lastly generate predictions using the conversation embedding. Conversation turns are frequently mapped to a vector embedding using CNN, RNN, and transformers (e.g. BERT), and conversation embeddings are derived from the turn embeddings using approaches like self-attention, RNN, Markov models, and graphical models (Mensio et al., 2018; Zayats and Ostendorf, 2018; Vickneswaran et al., 2020; Aliannejadi et al., 2020; Zeng et al., 2021). Drawing inspiration from these hierarchical approaches, we experiment with an approach where each turn is encoded using BERT and then the sequence of turn encodings is mapped to a fixed length vector using a uni-directional RNN. There is also conversation modeling work that encodes multiple turns as a single input sequence to BERT, separating the turns with the $[SEP]$ token (Huang et al., 2019), which we also explore here.

Lu et al. (2020) explored a retrieval-based response generation task and proposed a data augmentation technique for model training. The authors created additional positive samples by sampling contiguous multi-turn excerpts from conversations and assuming the last turn is a correct response. Additional negative samples were created by sampling contiguous multi-turn excerpts, randomly removing intermediate turns, and assuming the last turn is an incorrect response. We adapt this turn masking approach to our cognitive distor-

[1] https://www.liwc.app/

tion task to create dynamic conversation context in training, as described in Section 3.2.

3 Methods

3.1 Data

This work utilized a corpus of text message exchanges between patients and therapists that was created as part of a randomized controlled trial that augmented routine care for people with serious mental illness using a text-message-based intervention (Ben-Zeev et al., 2020). The trial was conducted in the Midwest and Pacific Northwest regions of the United States between December 2017 and October 2109. In the intervention, patients participating in standard care engaged with trained clinicians in back-and-forth text-message conversations for 12-weeks. Patients attended an in-person baseline visit to establish rapport and initial goals. Subsequently, clinicians attempted to contact patients up to three times a day to provide support strategies, including reminders, psycho-education, cognitive challenges, self-monitoring prompts, and relaxation techniques. Interactions could be initiated by either patient or clinician each day, and messages could be sent consecutively by a single party in cases where no response was given. The text-message exchanges represent a new model of care that is asynchronous and continuous. The trial demonstrated that augmenting care with mobile texting is logistically feasible, acceptable to patients, safe for patients, and clinically promising. A full description of the trial, including intervention feasibility, acceptability, engagement, and clinical outcomes is available (Ben-Zeev et al., 2020). The randomized controlled trial was approved by the University of Washington's Institutional Review Board (IRB), and study participants provided informed consent. Here, we utilize the text message data for secondary analysis with patient and therapist identifiers removed. All data were stored on a secure server, with patient and clinician identifiers removed prior to annotation and analysis.

The corpus created by the text-message intervention includes messages from 39 patients and 9 therapists with 7,436 patient and 6,959 therapist text messages. The patients who contributed data to the current analysis all had diagnoses of either schizophrenia, schizoaffective disorder, major depressive disorder, or bipolar disorder. The patient demographics were 56% male (N=22), 49% White (N= 17), 29% Black (N=10), 17% multiracial (N=6), and 8% Hispanic/Latinx (N=3). The patients had a mean age of 45.4 (SD=11.1), 12.8 years of education (SD=2.4), and 2.8 lifetime psychiatric hospitalizations (SD=3.4). Patients had variable levels of engagement in the text-message intervention with the average number of client messages per day ranging from 0.3 messages/day to 12.5 messages/day. The average length for the patient and therapist text messages is 15.9 and 22.0 tokens, respectively.

The text message conversations were annotated by a doctoral-level licensed mental health counselor and a masters-level psychologist experienced in working with people with serious mental illness. The corpus is annotated for six cognitive distortion types:

- *Catastrophizing (C)* - Exaggerating or discounting the importance of an event.
- *Jumping to conclusions (J)* - Interpreting a situation without facts or evidence, including mind reading and fortune telling.
- *Mental filtering (M)* - Focusing on one detail of a situation exclusively while ignoring other relevant information.
- *Should statements (S)* - Motivating oneself with absolute expectations, for example should, must, or ought.
- *Overgeneralization (O)* - Extending a single occurrence or isolated incident as evidence of an ongoing or never-ending pattern.
- *Unspecified (U)* - Message included a type of distortion not included in the five categories above or was too incoherent to code specifically.

Table 1 presents example text messages for each distortion type. Annotators reviewed text-messages in the context of a full patient-clinician transcript before applying cognitive distortion annotations at the individual message level. The therapist messages were used to interpret the patient messages; however, no cognitive distortion labels were assigned to therapist messages.

A patient text message may be annotated for multiple cognitive distortions. An *any distortion* (A) label was assigned to each patient text message, indicating whether there is at least one distortion type (logical "or" of distortion types at the message-level). Table 2 presents the distribution of the distortion types. Almost a third of the patient messages include distorted thinking; however, most

Distortion	Example
Catastrophizing (C)	"I just feel so emotional right now right now everything going wrong."
Jumping to conclusions (J)	"My family thinks I have no talents."
Mental filter (M)	"I can't say I have anything to be grateful for"
Should statements (S)	"The team is stopping by so I feel like I have to have my shit together."
Overgeneralizing (O)	"Its always hard to depend on people."
Unspecified (U)	"I felt like bugs were crawling on me and thought I saw some but didn't"

Table 1: Example text messages for each cognitive distortion type.

of the individual distortion types are relatively infrequent, resulting in an imbalanced label distribution. Approximately 20% of the annotated corpus was doubly annotated to assess inter-rater agreement. The Kappa values for the distortion types are: A=0.53, C=0.44, J=0.53, M=0.33, S=0.39, O=0.46, and U=0.01. To facilitate comparison with prediction performance, the inter-rater agreement was assessed as an F1 score, where one of two annotators was assumed to be the ground truth. Table 4 presents the inter-rater agreements as F1 scores. Notably, the agreement for the *unspecified* (U) category is considerably lower than for other categories.

Distortion	Count	Frequency
A	2,145	29%
C	1,113	15%
J	610	8%
M	656	9%
O	268	4%
S	198	3%
U	420	6%

Table 2: Label distribution.

3.2 Distortion Classification

3.2.1 Classification Task

We interpret this cognitive distortion prediction task as a multi-label binary text classification task, where the distortion label set is $\mathcal{V} = \{A, C, J, M, O, S, U\}$. For a given distortion type v in $\mathcal{V}$, a value of 1 indicates the presence of the cognitive distortion type in the target message, m_i. We explore the role of conversation history (context) in assessing the presence of distorted thinking by including preceding messages ($m_{i-n}, \ldots m_{i-2}$, m_{i-1}) in modeling, where n indicates the number of context messages or preceding turns used.

3.2.2 Classifier Architectures

We identify cognitive distortions in patient messages using two BERT architectures, which are presented in Figure 1. The first architecture, *BERT-only*, consists of BERT with a linear output layer operating on the pooled output vector. *BERT-only* encodes each target message, m_i, and the context messages, $m_{i-n}, \ldots m_{i-1}$, as a single input sequence, where the messages (turns) are delineated by the $[SEP]$ token. The input messages are ordered chronologically, so the last message is the target message ($m_{i-n}, \ldots, m_{i-1}, m_i$). The linear output layer projects the pooled output vector for the multi-turn conversation to the number of distortion types (7). In the second architecture, *BERT+LSTM*, each message is separately encoded by BERT, and the pooled output vectors for the messages are sequentially encoded using a uni-directional Long Short-Term Memory (LSTM) RNN. A linear output layer operating on the last hidden state of the LSTM generates the distortion type predictions. For both architectures, a sigmoid activation function converts the label scores to probabilities.

We experimented with including speaker role information to differentiate patients and therapists, for example, "[CLS] [therapist] After seeing her how is you anxiety? [SEP] [patient] It's ok ..." We also experimented with including patient and therapist identifiers, for example, "[CLS] [fe2k] After seeing her how is you anxiety? [SEP] [l2kd] It's ok ...," where "fe2k" and "l2kd" are unique anonymized identifiers for patients and therapists. These approaches did not yield a meaningful performance improvement and are omitted.

3.2.3 Message Context

We explore the introduction of additional randomness in the context messages ($m_{i-n}, \ldots, m_{i-1}$) to create dynamic context during model training. We investigate four context representation approaches: *none*, *fixed*, *random length*, and *random mask*. The

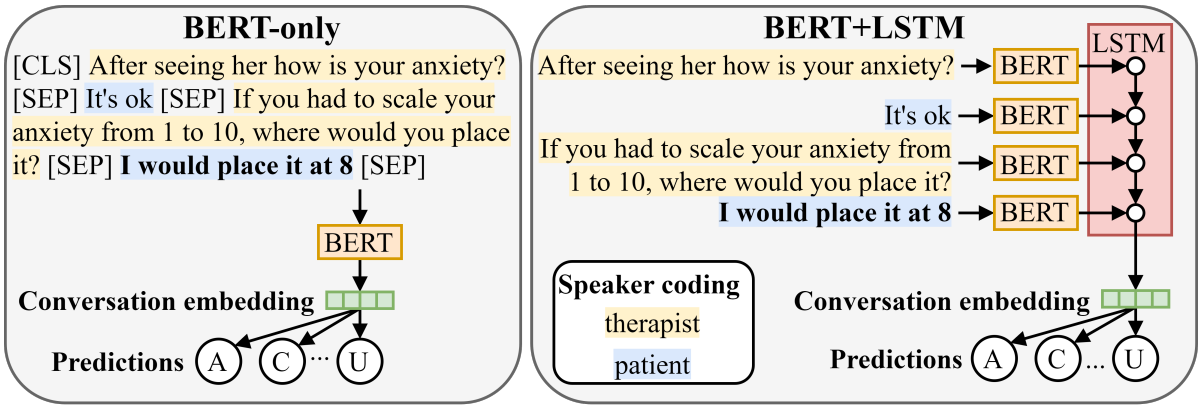

Figure 1: Conversation models. In the text-message examples, **boldface text** indicates the target message, and non-boldface text indicates the context messages.

none context approach does not incorporate any preceding messages as context ($n = 0$), and only the target message is used in training and inference. For the *fixed* context, n context messages preceding the target message are used in both training and inference. For the *random length* context, the number of context messages used in training is randomly selected from a uniform distribution ($uniform(0, n)$, inclusive) for each training sample. The *random length* approach provides contexts of varied lengths during training, and all context messages are sequential with the target message. For the *random mask* context, context messages are randomly masked (removed) for each training sample with probability, p_{mask}. Similar to *random length*, *random mask* provides target messages with varied context lengths; however, with *random mask* the context and target messages will not necessarily be contiguous, as some context messages are randomly removed. For the *random length* and *random mask* context approaches, n context messages are used in inference, similar to the *fixed* approach to utilize all available information. The context length, n, was treated as a tuneable hyperparameter, and context lengths from 0 to 4 were explored. Early experimentation demonstrated that prediction performance improves as the context length increases until $n = 3$, at which point the performance plateaus. All the presented results either include no context ($n = 0$ for *none*) or context of $n = 3$ for *fixed*, *random length*, and *random mask*.

3.2.4 Experimental Paradigm

Model performance was evaluated using a nested cross-validation procedure, to reduce error estimation bias (Varma and Simon, 2006). The annotated data set ($\mathcal{D}$) was split into five folds $(1, 2, ...5)$. To ensure each fold contains sequential messages, each patient-therapist conversation for the entirety of the study was arranged chronologically and split into five folds of approximately equal length ($\approx 20\%$ of each patient-therapist conversation in each fold). There was no overlap between the folds, such that a given message was only included as a target or context in a single fold. These folds were used to create train ($\mathcal{D}_{train}$), validation ($\mathcal{D}_{val}$), and test ($\mathcal{D}_{test}$) splits. Hyperparameters were tuned by training on $\mathcal{D}_{train}$ and evaluating on $\mathcal{D}_{val}$. Final model performance was assessed by training on $\mathcal{D}_{train} \cup \mathcal{D}_{val}$ and evaluating on $\mathcal{D}_{test}$. As a form of repeated holdout testing, we iterated over folds assigned to $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, and $\mathcal{D}_{test}$, retuning the hyperparameters for each iteration. For example, fold assignments for iteration $\#1$ were $\mathcal{D}_{train} = \{1, 2, 3\}$, $\mathcal{D}_{val} = 4$, $\mathcal{D}_{test} = 5$, iteration $\#2$ were $\mathcal{D}_{train} = \{2, 3, 4\}$, $\mathcal{D}_{val} = 5$, $\mathcal{D}_{test} = 1$, and so forth. Several of the distortions are very infrequent, and this nested cross validation procedure is intended to better characterize performance across the distortion types. Performance was averaged across the fold iterations and was assessed using F1-score. Hyperparameters were optimized to maximize average F1 across the fold iterations for the *any distortion* label. To assess final performance with significance testing, each fold iteration was repeated 10 times, to generate a distribution of 10 averaged F1 scores for each distortion type. Significance was evaluated using a two-sided T-test with unequal variance and a significance threshold of $p < 0.05$.

All presented results utilized the pretrained BERT model, *MentalBERT* (Ji et al., 2021), which was further pretrained on a Reddit corpus derived

Model	Context	Epochs by fold
BERT-only	none	[4, 4, 4, 6, 4]
BERT-only	all	[6, 4, 4, 4, 4]
	random length	[4, 4, 4, 4, 6]
	random mask	[6, 10, 4, 4, 8]
BERT+LSTM	all	[4, 8, 4, 4, 4]
	random length	[4, 4, 4, 6, 4]
	random mask	[4, 4, 6, 6, 4]

Table 3: Tuned hyperparameters by fold ($[1, 2, 3, 4, 5]$)

from mental health-related subreddits. Other pretrained models may offer performance gains over MentalBERT (Naseem et al., 2022); however, we leave this experimentation to future work. The following configuration and parameters were common to all experimentation: optimizer = AdamW, maximum gradient norm = 1.0, learning rate = 5e-5, batch size = 20, BERT dropout = 0.2, and maximum message length = 120 word pieces. For *BERT-only*, the maximum conversation length was 512 word pieces. For *BERT+LSTM*, the LSTM hidden size = 768. We experimented with context message counts, n, ranging from 0 to 4. We found that performance plateaus around $n = 3$. In the *random mask* experimentation, $p_{mask} = 0.2$. The number of training epochs was tuned for each fold and model configuration, and Table 3 presents the selected epochs for each configuration. To account for the class imbalance associated with label infrequency, a balanced loss function was used in all experimentation, where the loss weights for each label are inversely proportional to positive class frequency.

3.2.5 Distortion clustering

To explore the clustering of distortions in time and the role of conversation context, we calculated the pointwise mutual information (PMI) and conditional probability of distortions in the target and context messages. PMI assesses the association between events. To understand the relationship between distortions in the target message and preceding context messages, PMI is defined here as,

$$PMI(x = v, y = A) = log \frac{p(x = v, y = A)}{p(x = v)p(y = A)},$$

where x is the occurrence of distortion type $v \in \mathcal{V}$ in the target message, and y is the occurrence of *any distortion* (A) in the preceding context messages. We also assessed the association between distortions in target and context messages using the conditional probability, $P(y = A|x = v)$, where x and y are defined similarly to the PMI calculation.

4 Results

4.1 Prediction Performance

Table 4 presents the average cognitive distortion classification performance, as F1, averaged across 10 runs for each of the five fold iterations (each F1 score in the table is the average of 50 values). Each fold iteration involves training on the training and validation folds and evaluating on the withheld test fold. The *BERT-only* model with *none* context is the baseline model for evaluating the role of conversation history on prediction performance.

The inclusion of conversation context in the *BERT-only* and *BERT+LSTM* architectures yields an improvement over *BERT-only* without conversation context for a majority of the distortion labels. The *BERT-only* model with *random length* context achieved the best performance, with significance, for *any distortion* (A) and *catastrophizing* (C). The *BERT-only* model with *random mask* context achieved the best performance, with significance, for *jumping to conclusions* (J). The *BERT+LSTM* model with *fixed* context achieved the best performance, with significance, for *unspecified* (U). For the remaining distortion types (*mental filter* (M), *overgeneralizing* (O), and *should statements* (S)) there is not a statistically significant difference between the top performing model configurations. The dynamic context approaches, *random length* and *random mask*, yield a modest but statistically significant improvement over the *fixed* context for the more frequent and higher performing distortions (*any distortion*, *catastrophizing*, and *jumping to conclusions*).

4.2 Error Analysis

The results in Table 4 demonstrate the inclusion of preceding messages as context improves cognitive distortion prediction performance for the most frequently occurring distortions. We assessed the relationship between distortions in the target message and distortions in the context messages using the PMI, $PMI(x = v, y = A)$, and conditional probability, $P(y = A|x = v)$, defined in Section 3.2.5. Table 5 presents the PMI and conditional probabilities for the two data partitions, *All* and *Improved*. The *All* partition include all 7,436 pa-

Model	Context	F1						
		A (mean±STD)	**C**	**J**	**M**	**O**	**S**	**U**
BERT-only	none	0.68 ± 0.005	0.43	0.46	0.37	0.29	**0.20**	0.32
BERT-only	fixed	0.72 ± 0.003	0.47	0.47	**0.38**	0.29	0.19	0.31
	random length	**0.73** ± 0.003†	**0.48**†	0.46	0.37	0.29	**0.20**	0.33
	random mask	0.72 ± 0.003	0.46	**0.48**†	**0.38**	**0.30**	**0.20**	0.34
BERT+LSTM	fixed	0.72 ± 0.004	0.46	0.46	0.37	0.28	0.16	**0.38**†
	random length	0.72 ± 0.003	0.45	0.44	0.36	0.27	0.15	0.34
	random mask	0.72 ± 0.004	0.46	0.45	0.36	0.28	0.14	0.35
Inter-rater agreement		0.65	0.52	0.56	0.39	0.41	0.48	0.02

Table 4: Cognitive distortion prediction performance, averaged across 10 runs for each fold (1-5). The highest performance for each distortion is **bolded**, and † indicates the best performance with significance ($p < 0.05$). Performance for *any distortion* (A) is presented as mean ± standard deviation. Performance for the remaining distortion types is only presented as the mean, due to space constraints.

Inter-rater agreement is also presented for the doubly annotated subset of the corpus.

Distortion (v)	$PMI(x = v, y = A)$			$P(y = A \mid x = v)$		
	All	**Improved**	Δ	**All**	**Improved**	Δ
A	0.90	3.25	2.35	0.77	0.87	0.10
C	0.98	3.28	2.30	0.83	0.90	0.07
J	0.89	3.22	2.33	0.76	0.84	0.08
M	0.71	3.14	2.43	0.64	0.78	0.14
O	0.79	3.13	2.34	0.69	0.77	0.09
S	0.86	3.16	2.30	0.74	0.79	0.06
U	1.05	3.32	2.27	0.90	0.93	0.04

Table 5: PMI, $PMI(x = v, y = A)$ and conditional probability, $P(y = A|x = v)$, where x is the occurrence of distortion type v in the target message, and y is the occurrence of *any distortion* in the context messages.

tient messages in the annotated corpus. The PMI and conditional probability for *All* messages indicates that distortions cluster in time, specifically that distortions are more likely to occur in the context messages, if there are distortions in the target message (the reverse is also true).

We hypothesized that some of the improved distortion prediction performance associated with the inclusion of context is related with the model implicitly identifying distortions in the context messages. For *BERT-only* with *none* context and *BERT-only* with *random length* context, we identified the models that achieved median *any distortion* F1 performance amongst the 10 runs. We then identified all the samples for which the model without context (*BERT-only* with *none*) was incorrect in assigning the *any distortion* label and the model with context (*BERT-only* with *random length*) was correct is assigning the *any distortion* label. The *Improved* subset in Table 5 includes only the target messages where the model without context was incorrect and the model with context was correct in assigning the *any distortion* label. The *Improved* subset includes 535 target messages. In Table 5, the Δ columns indicates the change from *All* to *Improved*. The PMI and conditional probability are higher for the *Improved* partition across all distortion types, suggesting that at least a portion of the performance improvement associated with the inclusion of context is associated with the presence of distorted thinking in the context. The distortion types with the highest conditional probability in the *Improved* subset in Table 5 (*A*, *C*, *J*, and *U*) are also the distortion types for which the inclusion of context yielded a statistically significant improvement in prediction performance in Table 4.

#	Index	Speaker	Message
1	m_{i-3}	patient	my dad just recently has been trying to get to know me
	m_{i-2}	patient	I'm gonna call [NAME] but the voices r saying no
	m_{i-1}	therapist	Have the voices ever turned out wrong on what they said or ... told you to do?
	m_i	patient	**Some times they are**
2	m_{i-3}	therapist	I'd like to talk about what makes you nervous about leaving your house alone
	m_{i-2}	patient	I guess it started when I never left the house for all those years
	m_{i-1}	therapist	right. and what prevented you from leaving your house back then?
	m_i	patient	**I've never lived here before**

Table 6: Examples where the inclusion of context improves the performance for *any distortion*. In the text-message examples, **boldface text** indicates the target message, and non-boldface text indicates the context messages.

The *Improved* subset in Table 5 includes messages that were labeled incorrectly without the inclusion of context messages but labeled correctly when preceding messages were included as context. We manually reviewed the messages in this *Improved* subset to identify themes in the target and context messages. Table 6 presents example conversations that highlight two of the common themes identified during the manual review of the *Improved* subset. The examples in Table 6 were false negatives for the model without context and true positives for the model with context. In example $\#1$, the target message (m_i) is ambiguous and has no discernible meaning without context. With the inclusion of the context messages ($m_{i-3}, ...m_{i-1}$), we can infer that "they" refers to auditory hallucinations (voices) and "are" affirms that the voices are sometimes incorrect. There are many messages in the *Improved* subset, where the context messages confer meaning to otherwise ambiguous target messages. In example $\#2$, the target message has interpretable meaning without the preceding messages as context and does not necessarily convey distorted thinking. However, the preceding context messages include distorted thinking by the patient and a description of anxiety by the therapist. This context informs the interpretation of the target message and indicates the target message is a continuation of this distorted thinking. There are many examples where an individual message does not necessarily convey distorted thinking when viewed in isolation, but the broader context of the conversation indicates distorted thinking.

5 Discussion

We explored the automatic identification of cognitive distortions in text-based exchanges between patients and therapists, focusing on the role of conversation context. We utilized multiple transformer-based classification architectures and proposed two methods for dynamically utilizing conversation context in training, *random length* and *random mask*. Our results demonstrate that the inclusion of context improves cognitive distortion prediction performance for several distortion types, with the best performing architecture encoding the target message and context messages as a single input sequence to BERT (*BERT-only*). Results also demonstrate that using *random length* for the context during training improves performance over using a *fixed* length context, for several distortion types. The performance of the context-aware models approaches the inter-rater agreement for a majority of the distortion types. *BERT-only* with *random length* context identifies *any distortion* with relatively high performance at 0.73 F1; however, lower performance (F1 < 0.5) is achieved in resolving specific distortion types (e.g. *catastrophizing* or *jumping to conclusions*). The error analysis suggests that at least a portion of the performance gains associated with the inclusion of context messages is attributable to the tendency for messages expressing cognitive distortions to cluster in time.

This work presents context-aware classification approaches that improve performance in identifying cognitive distortions in text messages. The improved performance associated with the inclusion of context will benefit downstream clinical applications, including clinical decision-support systems, therapist training, and clinical research. In the community health setting, the adoption of new treatment modalities and technology for serious mental illness is hindered by the availability of training and expertise among community-based

clinicians (Perry et al., 2020). The adoption of new interventions is resource intensive, and training and supervision for novel interventions may improve the adoption of new interventions, like texting (Moyers et al., 2005). Our work exploring the automatic identification of cognitive distortions could mediate the development of clinician training and support tools that improve the uniformity and quality of care and reduce required human resources, by flagging patient content that requires intervention. In terms of clinical research, this work may support the implementation of interventions that target cognitive distortions, assess the extent to which such interventions are effective in reducing distortion frequency, and improve understanding of the relationships between distorted thinking, symptom severity and mental status.

This study is limited by the number of participating patients and therapists. Text-based therapy conversations are likely heterogeneous and vary by patient-therapist dyad, patient clinical condition, and other factors. Due to the size of the annotated corpus, the data set was split such that each patient appears both in the train and test partitions, although there is no overlap between the messages in the train and test partitions. Additional work with an expanded data set is needed to assess the generalizability of the classifiers to a diverse patient population, including patients not represented in the training data.

Similar to prior cognitive distortion work (Shickel et al., 2020), classification performance is limited by the challenge of manually annotating distortions, including the soft boundaries between distortion types. We are currently adding additional cognitive distortion type labels to the text-message corpus to include more fine-grained distortion categories that can be condensed into functionally related higher-level categories. The inclusion of additional cognitive distortion types and aggregation of individual distortion types into higher-level thought patterns may improve annotation consistency. As part of this annotation effort, we are expanding the annotation guidelines and providing additional annotator training to improve annotation detail and quality.

This work investigates the use of preceding conversational turns as context for prediction. There are many other forms of context, and mechanisms for representing it, that may be considered in future work. With a sufficiently large corpus of text conversations, it may be feasible to learn patient representations that capture important linguistic patterns, thinking styles, and other information relevant to characterizing thought patterns and mental state. The patient representations could take the form of learned patient embeddings, for example special patient-specific BERT tokens. Additional contextual information could include message metadata (e.g. time of day or time between responses) or patient demographics/attributes (e.g. age, gender, tech literacy, or diagnoses). Models incorporating such information may add to our understanding of the contexts in which distortions occur and further improve automated methods to detect them.

6 Conclusions

The improvements in performance shown in this work demonstrate that modeling conversational context is important for identifying cognitive distortions in text-based exchanges between patients and therapists. By identifying cognitive distortions in patient messages within the larger context of the conversation, the modeling better emulates the process mental health clinicians use to assess for distortions. Distorted thinking in the patient messages tends to cluster in time, such that distortions are more likely to occur in context messages, if there are distortions in the target message (and vice-versa). Some of the improved performance associated with the inclusion of context is likely attributable to the model implicitly identifying distortions in the context messages. Additionally, the inclusion of context also captures important cues in therapist messages for the presence of distorted thinking in patient messages. Conversational context is likely to improve performance in identifying cognitive distortions, with implications for the development of decision support tools, and quantification of distortions in observational data.

Acknowledgements

This work was supported by a UW Medicine Garvey Institute for Brain Health Solutions Innovation Grant, a grant from the National Institute of Mental Health (R56MH109554), and the National Institutes of Health, National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at UW (Grant Nr. T15LM007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *Proceedings of the Conference on Human Information Interaction and Retrieval*, page 33–42.

Aaron T Beck. 1963. Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, 9(4):324–333.

Dror Ben-Zeev, Benjamin Buck, Suzanne Meller, William J Hudenko, and Kevin A Hallgren. 2020. Augmenting evidence-based care with a texting mobile interventionist: a pilot randomized controlled trial. *Psychiatric Services*, 71(12):1218–1224.

David D Burns. 1980. *Feeling Good: The New Mood Therapy*. William Morrow and Company.

Jessica D'Arcey, Joanna Collaton, Nicole Kozloff, Aristotle N Voineskos, Sean A Kidd, George Foussias, et al. 2020. The use of text messaging to improve clinical engagement for individuals with psychosis: systematic review. *Journal of Medical Internet Research - Mental Health*, 7(4):e16993.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Robert Dudley, Peter Taylor, Sophie Wickham, and Paul Hutton. 2016. Psychosis, delusions and the "jumping to conclusions" reasoning bias: a systematic review and meta-analysis. *Schizophrenia Bulletin*, 42(3):652–665.

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, 36(5):427–440.

Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019. EmotionX-IDEA: Emotion BERT–an affectional model for conversation. *arXiv preprint*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly available pretrained language models for mental healthcare. *arXiv preprint*.

Amanda Lin and Alberto J Espay. 2021. Remote delivery of cognitive behavioral therapy to patients with functional neurological disorders: Promise and challenges. *Epilepsy & Behavior Reports*, 16:100469.

Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. 2020. Improving contextual language models for response retrieval in multi-turn conversation. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1805–1808.

Martino Mensio, Giuseppe Rizzo, and Maurizio Morisio. 2018. Multi-turn QA: A RNN contextual approach to intent classification for goal-oriented systems. In *International World Wide Web Conference - Companion Proceedings*, pages 1075–1080.

Theresa B. Moyers, Tim Martin, Jennifer K. Manuel, Stacey M.L. Hendrickson, and William R. Miller. 2005. Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28.

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam Dunn. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 22–31. Association for Computational Linguistics.

Kristen Perry, Sari Gold, and Erika M. Shearer. 2020. Identifying and addressing mental health providers' perceived barriers to clinical video telehealth utilization. *Journal of Clinical Psychology*, 76(6):1125–1134.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *IEEE International Conference on Bioinformatics and Bioengineering*, pages 275–280.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158.

Taetem Simms, Clayton Ramstedt, Megan Rich, Michael Richards, Tony Martinez, and Christophe Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *IEEE International Conference on Healthcare Informatics*, pages 508–512.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):1–8.

Jarsigan Vickneswaran, Piruntha Navanesan, Vahesan Vijayaratnam, and Uthayasanker Thayasivam. 2020. Simplified approach for predicting emotions of multi-turn textual utterances. In *International Conference on Advances in ICT for Emerging Regions*, pages 71–76.

Katja Wiemer-Hastings, Adrian S Janit, Peter M Wiemer-Hastings, Steve Cromer, and Jennifer Kinser. 2004. Automatic classification of dysfunctional thoughts: a feasibility test. *Behavior Research Methods, Instruments, & Computers*, 36(2):203–212.

Carrie L. Yurica and Robert A. DiTomasso. 2005. *Cognitive distortions*, pages 117–122. Springer.

Victoria Zayats and Mari Ostendorf. 2018. Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics*, 6:121–132.

Xingshan Zeng, Jing Li, Lingzhi Wang, and Kam-Fai Wong. 2021. Modeling global and local interactions for online conversation recommendation. *ACM Transactions on Information Systems*, 40(3):1–33.