Automatic Detection and Classification of Cognitive Distortions in Mental Health Text

Benjamin Shickel*§, Scott Siegel†§, Martin Heesacker‡§, Sherry Benton¶, Parisa Rashidi†§
*Department of Computer and Information Science and Engineering, †Department of Biomedical Engineering,
‡Department of Psychology, §University of Florida, ¶TAO Connect

***Abstract*—In cognitive psychology, automatic and self-reinforcing irrational thought patterns are known as cognitive distortions. Left unchecked, patients exhibiting these types of thoughts can become stuck in negative feedback loops of unhealthy thinking, leading to inaccurate perceptions of reality commonly associated with anxiety and depression. In this paper, we present a machine learning framework for the automatic detection and classification of 15 common cognitive distortions in two novel mental health free datasets collected from both crowdsourcing and a real-world online therapy program. We also performed an exploratory analysis using unsupervised content-based clustering and topic modeling algorithms as first efforts towards a data-driven perspective on the thematic relationship between similar cognitive distortions traditionally deemed unique. Finally, we highlight the difficulties in applying mental health-based machine learning in a real-world setting and comment on the implications and benefits of our framework for improving automated delivery of therapeutic treatment in conjunction with traditional cognitive-behavioral therapy.**

I. Introduction

Cognitive-behavioral therapy (CBT) has been shown to have a large positive effect on patients experiencing symptoms of anxiety and depression [1]. A large part of CBT treatment is helping the patient learn to self-identify their automatic and often irrational thought patterns that contribute to a distorted perception of reality. These patterns are known as cognitive distortions.

From a cognitive-behavioral perspective, cognitions can be viewed as the intermediary link between an external stimulus and a subjective feeling, taking the form of internalized statements a person tells themselves [2]. Cognitive distortions denote such self-statements that are either mildly misinterpreted or entirely inaccurate, often reflected by logical fallacies in internal thinking. For example, *"they didn't respond to my text message, they must be ignoring me"* is a cognitive distortion because it makes definitive assumptions about a third party that is impossible to know with certainty. In this example, a non-distorted cognition might be *"they didn't respond to my text message, maybe they are preoccupied or haven't seen it yet"*. In many cases, negative subjective feelings such as anxiety are responses to these types of distorted cognitions. A large focus of CBT involves training patients to identify their own cognitive misjudgments to improve overall well-being.

In recent years, online therapy programs have been developed to supplement or replace traditional cognitive-behavioral therapy [3]–[5]. Similar to traditional CBT, these online counterparts provide guidance and instruction for identifying and challenging cognitive distortions and negative feedback loops. A frequent component of these online services is a self-monitoring journal, where patients describe their thoughts, emotions, anxiety events, and self-assessment.

In this paper, we develop methods for automatically detecting and classifying cognitive distortions in mental health text using machine learning techniques. Such frameworks have important implications for the delivery of effective online mental health services. Along with assisting traditional therapist intervention, automated distortion assessment can provide instant feedback and guidance to users exhibiting distorted thinking at any time of day. These frameworks would also allow for online therapy to scale to an even larger number of users. Additionally, these tools can provide early warning of more severe mental illness and can potentially be used to monitor social media for ideal candidates for mental health services.

While machine learning and deep learning methods have been successfully applied to many natural language processing classification tasks in other domains, relatively few works have explored these techniques in the context of mental health. Within this space, researchers have focused on tasks including but not limited to emotional valence prediction in mental health journals [6], identifying characteristic language indicators of mental health in social media [7]–[9], sentiment analysis of suicide notes [10], and early warning indicators for poor mental health in online forums [11]. Less attention has been paid to applying machine learning techniques from a cognitive-behavioral perspective.

Cognitive distortion detection and classification share similarities with the task of text-based emotion recognition [12]–[14]. To our knowledge, only [13] has explicitly dealt with the classification of cognitive distortions. However, [13] used a list of five distortions, compared with our expanded list of 15 [15]. Furthermore, they worked exclusively with crowdsourced data. To our knowledge we are the first to classify cognitive distortions using real-world, unscripted mental health text from online therapy programs.

Our primary contributions can be summarized by the following:

- To our knowledge, we present the first attempts at detecting and classifying a large range of cognitive distortions (15) from text using machine learning techniques.
- We collect a novel dataset of real-world cognitive distortion event recollections from crowdsourcing participants.

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00052

- We develop a second novel dataset of online mental health therapy logs annotated by experts for presence of cognitive distortions.
- We examine thematic similarities and overlap among recognized cognitive distortions based on unsupervised machine learning techniques.

II. Methods

A. Data

Since cognitive distortion detection is a novel machine learning task, at present time there are no publicly available datasets containing text passages with labeled distortions. Such data are necessary for constructing supervised machine learning models suitable for predicting and classifying text-based distortions. To this end, we collected and annotated three novel cognitive distortion datasets, which we describe in this section. These datasets are summarized in Table. I. The CrowdDist and MH-C datasets are labeled with 15 cognitive distortions, while MH-D is annotated with binary distortion/no distortion labels.

1) Crowdsourced distortion recollections (CrowdDist): Our first cognitive distortions dataset, which we refer to as CrowdDist, comes from the popular crowdsourcing platform Mechanical Turk[1] (MTurk). MTurk workers were presented with a short description of the types of thinking that a particular distortion exemplifies, and were asked to describe a specific time from their own life in which they exhibited the same type of distorted thinking. Distortions were randomly assigned to arrive at an even distribution of labeled responses.

In total, we collected 8,940 text passages from 1,788 unique individuals from MTurk's pool of qualified workers. After manually removing irrelevant contributions, the final dataset contained 7,666 text responses across all 15 distortions, with an average of 511 responses per distortion.

2) Mental health therapy logs (MH): Our second dataset, which we refer to as MH, comes from TAO Connect[2], an online mental health therapy service aimed at improving coping strategies for college students suffering from anxiety and depression. As part of treatment, patients completed regular journals describing their mental state, anxiety events, specific worries, and progress. We recruited four senior psychology students to provide cognitive distortion annotations for these mental health journals. Each annotator was instructed to select one or more cognitive distortions that each journal response exhibited, or to indicate that no distortions were exhibited.

From the overall MH dataset, we constructed two subsets for each of our two cognitive distortion tasks. The MH-C dataset was annotated with 15 cognitive distortion labels, and the MH-D dataset was annotated with binary distorted/non-distorted labels. The MH-C dataset was used in the 15-class cognitive distortion classification task. We only kept passages in which a majority of annotators agreed on which cognitive distortion was exemplified by the passage. For detecting presence of distorted text, we assigned a binary label to each passage indicating whether any cognitive distortion was selected by the annotators.

3) Preprocessing: For both cognitive distortion tasks, we applied a sequence of traditional techniques for processing natural language text before passing to our machine learning models for prediction, including lowercasing, punctuation removal, tokenization, and extraction of unigrams and bigrams using term frequency-inverse document frequency (tf-idf).

B. Model

For both classification tasks, we experimented with a variety of machine learning models including logistic regression, support vector machines (SVM), random forests, gradient boosted trees (XGBoost), recurrent neural networks (RNN), convolutional neural networks (CNN), and the recent Bidirectional Encoder Representations from Transformers (BERT) [16].

For simplicity, we only report results for the best-performing model: logistic regression. In contrast with more recent deep learning models in which interpretability is often sacrificed for predictive performance, logistic regression has the added benefit of being highly transparent, and in our datasets also resulted in superlative performance compared with more complex approaches. We hypothesize that logistic regression outperformed deep learning techniques for our particular tasks due to the relatively small size of our datasets, and the tendency of cognitive distortions to present themselves via specific, common emotional expressions that bag-of-words-based models more easily capture. We note in passing that aside from logistic regression, the second-best model for both cognitive distortion tasks was BERT, the recent technique for transferring unsupervised language models to downstream NLP tasks. We refer interested readers to [16] for a more comprehensive overview of BERT and modern language model transfer techniques.

C. Classification experiments

We define cognitive distortion detection as the ability for a machine learning classifier to distinguish between text containing a distortion and text that is not distorted, and cognitive distortion classification as the ability for a model to distinguish between the 15 distortion categories in text that is already known to be distorted. Both tasks are important from a mental health perspective, as once text is determined to contain some distortion, knowing which distortion is present can guide treatment options for that patient. Because our CrowdDist dataset exclusively contains distorted text, we only report its distortion classification results.

When evaluating each individual dataset, we report the 5-fold nested cross-validation performance. When possible, we stratified the folds based on the distribution of distortions occurring in the dataset. In the MH-C dataset obtained from actual patients, some distortions resulted in very few annotated instances, so these instances were assigned to folds randomly.

D. Unsupervised exploration

In this study, we used the set of 15 cognitive distortions outlined in recent literature [17]; however, there is still no clear

[1] http://mturk.com

[2] http://taoconnect.org

TABLE I
SUMMARY STATISTICS FOR CROWDDIST AND MH DATASETS.

	CrowdDist	**MH-C**	**MH-D**
Task	Classification	Classification	Detection
Source	Mechanical Turk	Mental health logs	Mental health logs
Guided collection	Yes	No	No
Passages, n	7,666	1,164	1,799
Unique labels, n	15	15	2
Median passage length, words	47	43	42
Annotators per passage, n	1	4	4
Distortion prevalence, n (%)			
Distorted	7,666 (100.0)	1,164 (100.0)	1,605 (89.2)
Not Distorted	0 (0.0)	0 (0.0)	194 (10.8)
Distortion counts, n (%)			
Being Right	536 (7.0)	0 (0.0)	—
Blaming	494 (6.4)	23 (2.0)	—
Catastrophizing	545 (7.1)	53 (4.6)	—
Control Fallacy	490 (6.4)	60 (5.2)	—
Emotional Reasoning	500 (6.5)	187 (16.1)	—
Fallacy of Change	499 (6.5)	0 (0.0)	—
Fallacy of Fairness	495 (6.5)	1 (0.1)	—
Filtering	545 (7.1)	386 (33.2)	—
Global Labeling	493 (6.4)	4 (0.3)	—
Heaven's Reward Fallacy	490 (6.4)	0 (0.0)	—
Mind Reading	545 (7.1)	260 (22.3)	—
Overgeneralization	546 (7.1)	22 (1.9)	—
Personalization	497 (6.5)	25 (2.1)	—
Polarized Thinking	497 (6.5)	123 (10.6)	—
Should's	494 (6.4)	20 (1.7)	—

evidence-based consensus in the mental health and psychology communities on the best way to classify cognitive distortions. Various studies argue for lists ranging from five [13] to 50 [18] cognitive distortions. In this section, we detail our unsupervised exploratory analysis procedure using the text in both of our cognitive distortion datasets. For all unsupervised analysis, we use the tf-idf representations of each passage as feature vectors.

1) Hierarchical clustering: We examined both datasets from a hierarchical clustering perspective, with the hypothesis that there may exist natural and hierarchical groupings of cognitive distortions that share common traits. For improved visualization, we took the sum of all tf-idf representations for each of the 15 cognitive distortions, resulting in a single aggregate feature vector per distortion. Clustering was performed with Ward's method of decreasing cluster variance using cosine similarity as the measure of distance.

2) Latent Dirichlet allocation: We also used Latent Dirichlet Allocation (LDA), a popular unsupervised topic-modeling algorithm for extracting a set of thematic topics composed of distributions of words from a corpus of text documents. It is a probabilistic method which assumes each document is generated by sampling words from a distribution of topics, where each topic is a distribution over the words in the corpus. In our experiments, we found n = 25 topics to yield the most meaningful topics, and we use this setting for LDA-based analysis. Once the LDA model was trained on an entire corpus, each document was converted into a 25-dimensional probability distribution over topics. For each of the 15 cognitive distortions, we took the sum of all topic probabilities from each of the distortion's passages as the overall LDA representation of the given distortion. We compared the pairwise cosine similarity between each pair of distortions as a measure of how related two distortions were in terms of their thematic content, repeating this process for every pair of two cognitive distortions.

III. RESULTS

In this section, we begin by reporting classification performance for both the cognitive distortion detection and classification tasks. We then describe the results of our unsupervised distortion analysis.

A. Cognitive distortion detection

Our first task, in which we wish to predict whether a text passage contains evidence of a cognitive distortion or is non-distorted, only applies to our online therapy dataset. In the CrowdDist dataset, in which we directly elicited responses from participants matching each of the 15 cognitive distortions, we do not have text that is labeled as non-distorted.

The 5-fold cross validation classification experiments resulted in a weighted F1 score of 0.88 for all classes (Table. II), with the weighted precision, recall, and F1 score for the distorted class (0.92, 0.97, 0.95, respectively) greatly outperforming the same non-distorted classification metrics (0.57, 0.29, 0.38, respectively). Accuracy for this task was 0.90, but given the significant class imbalance, we place more emphasis on F1 score for our primary classification metric.

TABLE II
COGNITIVE DISTORTION DETECTION RESULTS IN THE MH-D DATASET.

Label	N	Precision	Recall	F1
Not Distorted	194	0.57	0.29	0.38
Distorted	1605	0.92	0.97	0.95
All Examples (Macro)	**1799**	**0.75**	**0.63**	**0.66**
All Examples (Weighted)	**1799**	**0.88**	**0.90**	**0.88**

B. Cognitive distortion classification

Our second task involved the classification of distorted text into the 15 distinct distortion categories shown in Table. I. For this task, each response is known to contain some unspecified cognitive distortion, whether by crowdsourced volunteers (CrowdDist) or annotated as such from online mental health journals (MH-C). Our aim was to predict which distortion the text contains. Specific, fine-grained identification of cognitive distortions is important for tailoring mental health treatment and providing specific feedback to both patients and therapists.

We note that the distribution of distortion labels is roughly balanced in the CrowdDist dataset (Table. I). This is a direct result of the guided nature of our data collection, where we intentionally collected a balanced distribution of responses for each of the 15 cognitive distortion categories. In contrast, the MH-C dataset comes from real-world patient journals in a currently active online mental health therapy service. The distribution of distortions in MH-C is much different, with some labels receiving zero annotator votes. While CrowdDist can be viewed as the ideal classification setting, MH-C exhibits the difficulties that come with modeling real-world data.

Table. III shows the precision, recall, and F1 scores for the cognitive distortion classification task in both datasets. The CrowdDist model yielded 0.68 accuracy and the MH-C model yielded 0.47 accuracy. Given 15 possible distortion labels, the accuracy represented by random chance is 0.06.

C. Unsupervised exploration

Aside from constructing models to accurately predict and classify distorted text passages, we also sought to quantitatively suggest the natural number of unique distortions in a data-driven manner. For these experiments, we did not assume any fixed number of cognitive distortions, which is frequently debated in literature. Instead, we turned to unsupervised machine learning techniques to reveal natural co-occurrence and groupings of related text in both CrowdDist and MH-C passages containing cognitive distortions. For brevity, in this section we only show unsupervised analysis of the CrowdDist dataset, which is larger and more balanced than the MH dataset.

The result of hierarchical clustering on the distorted text passages from the CrowdDist dataset is shown as a dendrogram in Fig. 1. The clustering via Ward's method using cosine distance appears to suggest four natural groupings of cognitive distortions.

In Fig. 2, we examine the cosine similarity between every pair of our 15 cognitive distortions in the CrowdDist dataset, evaluated based on the sum of the LDA topic distribution for each passage of each cognitive distortion. In the CrowdDist dataset, the five most similar distortion pairs were Fallacy of Fairness/Heaven's Reward Fallacy (0.98), Emotional Reasoning/Mind Reading (0.98), Emotional Reasoning/Global Labeling (0.98), Being Right/Mind Reading (0.98), and Global Labeling/Mind Reading (0.95). In the MH-C dataset, the five most similar distortion pairs were Filtering/Polarized Thinking (0.98), Blaming/Filtering (0.97), Blaming/Emotional Reasoning (0.97), Catastrophizing/Filtering (0.96), and Blaming/Polarized Thinking (0.96).

IV. DISCUSSION

Our cognitive distortion detection model performed generally well with a weighted F1 score of 0.88 across all passages (Table. II). However, while the classifier was fully capable of correctly identifying distorted text (0.92 precision, 0.97 recall, 0.95 F1), it struggled with non-distorted passages from the same domain (0.57 precision, 0.29 recall, 0.38 F1 score). One rationale for poor performance in non-distorted text is the class imbalance in the MH-D dataset: only 10.8% of all text passages contained zero cognitive distortions. Since MH-D comes from a real-world mental health service, it is a logical expectation that most user journals will involve some type of distorted thinking. Furthermore, the relatively small size of MH-D (1,799 passages) would limit the ability of fully developing a domain-dependent language model and differentiating between text of each distortion class with very high accuracy.

When classifying the primary cognitive distortion of text already known to be distorted, our model performed well in the CrowdDist dataset with a weighted F1 score across all passages of 0.68 (Table. III), representing a clear improvement over the random-chance baseline of 0.06 given the 15 possible prediction classes. F1 scores for individual distortions ranged from 0.55 (Emotional Reasoning) to 0.77 (Fallacy of Fairness). While classification performance was overall satisfactory in the CrowdDist dataset, we believe that the short nature of the passages and the potential presence of multiple distortions had a negative impact on classification. Since distortions are known to co-occur and several distortions are similar in thematic content, our unsupervised analysis was designed to provide justification for reducing the number of cognitive distortions.

Distortion classification performance in the MH-C dataset was fair, with an overall weighted F1 score across all passages of 0.45 (Table. III), ranging from 0 (Blaming, Fallacy of Fairness, Global Labeling, Personalization) to 0.56 (Filtering). The primary difficulty in MH-C was the lack of annotation quantity for five distortions (Being Right, Fallacy of Change, Fallacy of Fairness, Global Labeling, Heaven's Reward Fallacy). Given the small dataset size and these underrepresented distortions, the 15-class classification problem proved especially difficult in the mental health domain. While psychologists and other mental health experts have identified 15 distinct cognitive distortions, our results indicate that some distortions occur much more frequently than others in a real-world setting. Combined with our unsupervised analysis of distortion similarity, we feel

TABLE III
COGNITIVE DISTORTION CLASSIFICATION RESULTS IN CROWDDIST AND MH-C DATASETS.

	CrowdDist				**MH-C**			
Label	**N**	**Precision**	**Recall**	**F1**	**N**	**Precision**	**Recall**	**F1**
Being Right	536	0.73	0.78	0.75	0	—	—	—
Blaming	494	0.77	0.72	0.74	23	0.00	0.00	0.00
Catastrophizing	545	0.67	0.71	0.69	53	0.50	0.21	0.29
Control Fallacy	490	0.65	0.64	0.64	60	0.28	0.22	0.25
Emotional Reasoning	500	0.55	0.55	0.55	187	0.40	0.40	0.40
Fallacy of Change	499	0.74	0.75	0.75	0	—	—	—
Fallacy of Fairness	495	0.81	0.73	0.77	1	0.00	0.00	0.00
Filtering	545	0.70	0.74	0.72	386	0.52	0.62	0.56
Global Labeling	493	0.61	0.53	0.57	4	0.00	0.00	0.00
Heaven's Reward Fallacy	490	0.65	0.67	0.66	0	—	—	—
Mind Reading	545	0.61	0.67	0.64	260	0.50	0.60	0.55
Overgeneralization	546	0.64	0.63	0.63	22	0.43	0.14	0.21
Personalization	497	0.68	0.63	0.66	25	0.00	0.00	0.00
Polarized Thinking	497	0.61	0.60	0.61	123	0.39	0.37	0.38
Should's	494	0.73	0.77	0.75	20	0.50	0.05	0.09
All Examples (Macro)	**7666**	**0.68**	**0.68**	**0.68**	**1164**	**0.29**	**0.22**	**0.23**
All Examples (Weighted)	**7666**	**0.68**	**0.68**	**0.68**	**1164**	**0.44**	**0.47**	**0.45**

it is important to revisit these distortion categories for possible adjustment based on real patient data.

A qualitative analysis of the hierarchical clustering experiments in the CrowdDist dataset (Fig. 1) would indicate four inherent clusters. There was no trend in the groupings of specific distortions between datasets, suggesting that thematic content in each domain played a large role in the way each distortion was exemplified. We compared the cosine similarity between topic distributions using latent Dirichlet allocation in an attempt to quantify the relatedness between each distortion pair's thematic content (Fig. 2). Each domain produced different thematic similarities, but here we note some patterns between domains; for example, in both datasets, Polarized Thinking is highly similar to Should's, and Emotional Reasoning is highly dissimilar to Fallacy of Change.

Above all, both supervised and unsupervised experiments involving cognitive distortions highlight the importance of source domain when dealing with free text passages. We have showed that distorted text takes on different forms when coming from third-party individuals (CrowdDist) or real patients currently experiencing distorted thinking (MH). These experiments indicate the nature of cognitive distortions themselves: what is clear to healthy individuals may be more difficult to understand for patients affected by certain types of distorted thinking.

Cognitive distortions put people at significant risk for engaging in and sustaining dysfunctional behaviors. A machine learning tool for detecting cognitive distortions from patient responses represents a critically important advance in computer-assisted healthcare delivery. This tool can be used to alert clinicians about the presence and frequency of distorted thinking exhibited in clients' verbal or written behavior, and to focus treatment toward changing distorted cognitions. One of the most useful features of this tool is that it is noninvasive; clients simply engage in the normal verbal or written behavior associated with treatment and the machine learning tool assesses for distortions without any additional intervention.

Accurately typing specific cognitive distortions and determining how many discrete distortions are present in a particular piece of client verbal behavior remain challenging. What we noticed in patient data that was absent in crowd-sourced data were cascades of cognitive distortions, where two or more cases of distorted thinking were exhibited in a single verbal response. Disentangling the pieces of these cascades is challenging. A closely related challenge involves improving the accuracy of distortion typology. Some scholars have identified 10 distinct distortions, others 15, and some others have provided other estimates. Comparison of these lists reveals both the expected overlap between lists, but also often overlap within lists. All of this suggests that in the future, factor analytic and other statistical techniques for grouping distortions need to be employed to provide empirical clarity regarding the number of independent (or interdependent) dimensions of distortion that most accurately characterizes clients' cognitively distorted verbal behavior.

Our study comes with several limitations. Our CrowdDist dataset was collected by prompting volunteers with specific descriptions of distortions, and it is possible that the types of responses we gathered based on these scripted prompts do not accurately represent the types of mental health text found "in the wild". Additionally, given the small size of our real-world MH dataset, it is difficult to draw definitive conclusions when not all distortions were equally represented in the data. Future work in this area should primarily focus on robust data collection.

V. CONCLUSION

This study details the application of machine learning techniques toward detecting, classifying, and understanding the underlying structure of cognitive distortions in short text passages. There is a currently a substantial lack of annotated

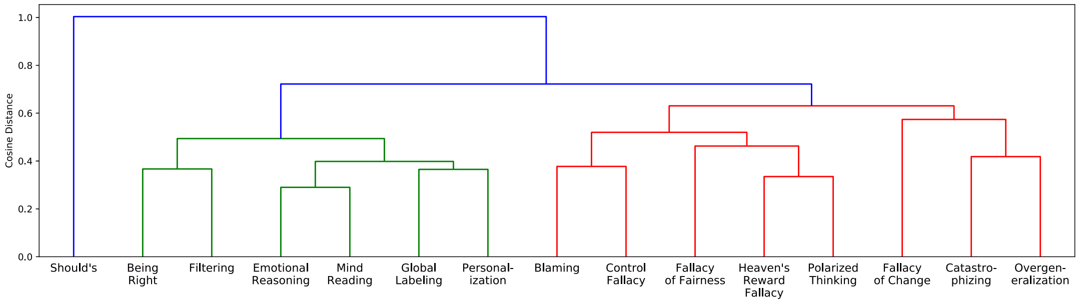

Fig. 1. Hierarchical clustering results for the CrowdDist dataset. Feature vectors for each of the 15 cognitive distortions were obtained by taking the sum of the tf-idf representations of passages for each label. Clusters obtained via Ward's method using cosine similarity as the measure of distance.

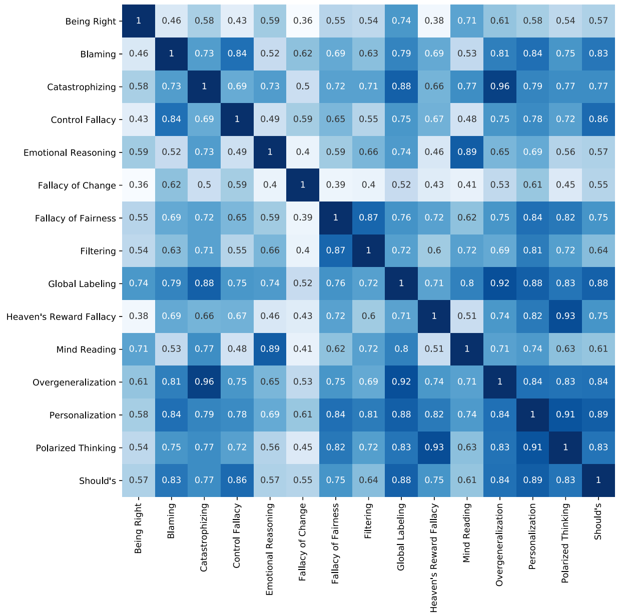

Fig. 2. Cosine similarity between the sum of each labeled document's 25-topic LDA topic distribution in the CrowdDist dataset. Pairs of cognitive distortions that are more similar in thematic content are shown as darker squares.

datasets in this domain, and one of our primary contributions is the collection of two novel cognitive distortion datasets coming from both crowdsourcing volunteers (CrowdDist) and real mental health patients (MH). Additionally, our unsupervised analysis provided multiple perspectives on distortion similarity that represent the first steps toward a data-driven rationale for revisiting the distinction between distortions, and for possible reduction in their overall number.

Acknowledgment

Research supported by NSF-IIP 1631871 from the National Science Foundation (NSF), Division of Industrial Innovation and Partnerships (IIP).

References

[1] S. G. Hofmann, A. Asnaani, I. J. J. Vonk, A. T. Sawyer, and A. Fang, "The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses."

[2] A. T. Beck, "Cognitive Therapy: Nature and Relation to Behavior Therapy," *Behavior Therapy*, vol. 1, no. 2, pp. 184–200, 1970.

[3] J. Ruwaard, A. Lange, B. Schrieken, C. V. Dolan, and P. Emmelkamp, "The effectiveness of online cognitive behavioral treatment in routine clinical practice," *PLoS ONE*, vol. 7, no. 7, 2012.

[4] M. F. Travers and S. A. Benton, "The Acceptability of Therapist-Assisted, Internet-Delivered Treatment for College Students," *Journal of College Student Psychotherapy*, vol. 28, no. 1, pp. 35–46, 2014.

[5] S. A. Benton, M. Heesacker, S. J. Snowden, and G. Lee, "Therapist-assisted, online (TAO) intervention for anxiety in college students: TAO outperformed treatment as usual." *Professional Psychology: Research and Practice*, vol. 47, no. 5, pp. 363–371, 2016.

[6] B. Shickel, M. Heesacker, S. Benton, A. Ebadi, P. Nickerson, and P. Rashidi, "Self-Reflective Sentiment Analysis," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 23–32.

[7] J. D. Hwang and K. Hollingshead, "Crazy Mad Nutters: The Language of Mental Health," in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, jun 2016, pp. 52–62.

[8] G. Gkotsis, A. Oellrich, T. Hubbard, R. Dobson, M. Liakata, S. Velupillai, and R. Dutta, "The language of mental health problems in social media," in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, jun 2016, pp. 63–73.

[9] M. Tanana, A. Dembe, C. S. Soma, Z. Imel, D. Atkins, and V. Srikumar, "Is Sentiment in Movies the Same as Sentiment in Psychotherapy? Comparisons Using a New Psychotherapy Sentiment Database," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, jun 2016, pp. 33–41.

[10] J. Pestian, P. Matykiewicz, Brett South, Ozlem Uzuner, and John Hurdle, "Sentiment analysis of suicide notes: A shared task," *Biomedical Informatics Insights*, vol. 5, no. 1, pp. 3–16, 2012.

[11] B. Shickel and P. Rashidi, "Automatic Triage of Mental Health Forum Posts," in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, 2016, pp. 188–192.

[12] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, p. 1556, 2008.

[13] R. Morris and R. Picard, "Crowdsourcing collective emotional intelligence," *Collective Intelligence Conference*, pp. 1–8, 2012.

[14] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[15] A. T. Beck, *Cognitive therapy and the emotional disorders*. New York: Meridian, 1976.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[17] J. M. Grohol, "15 Common Cognitive Distortions," 2018. [Online]. Available: https://psychcentral.com/lib/15-common-cognitive-distortions/

[18] A. Boyes, "50 Common Cognitive Distortions." [Online]. Available: https://www.psychologytoday.com/blog/in-practice/201301/50-common-cognitive-distortions