Improving Classification of Infrequent Cognitive Distortions: Domain-Specific Model vs. Data Augmentation

Xiruo Ding and **Kevin Lybarger** and **Justin S. Tauscher** and **Trevor Cohen**
{xiruod, lybarger, jtausch, cohenta}@uw.edu
University of Washington, Seattle, WA, USA

Abstract

Cognitive distortions are counterproductive patterns of thinking that are one of the targets of cognitive behavioral therapy (CBT). These can be challenging for clinicians to detect, especially those without extensive CBT training or supervision. Text classification methods can approximate expert clinician judgment in the detection of frequently occurring cognitive distortions in text-based therapy messages. However, performance with infrequent distortions is relatively poor. In this study, we address this sparsity problem with two approaches: *Data Augmentation* and *Domain-Specific Model*. The first approach includes Easy Data Augmentation, back translation, and `mixup` techniques. The second approach utilizes a domain-specific pretrained language model, MentalBERT. To examine the viability of different data augmentation methods, we utilized a real-world dataset of texts between therapists and clients diagnosed with serious mental illness that was annotated for distorted thinking. We found that with optimized parameter settings, `mixup` was helpful for rare classes. Performance improvements with an augmented model, MentalBERT, exceed those obtained with data augmentation.

1 Introduction

Data augmentation first became a popular topic in computer vision, where deep neural networks have performed remarkably well. Complex architectures, such as AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), generally require sufficient training data for model convergence, even with the help of dropout regularization and batch normalization. This situation also occurs in natural language processing (NLP) with deep learning methods and can become more problematic when limited to small datasets by data collection or data annotation constraints. In imaging, data augmentation, involving transformations such as cropping and shearing, is a common strategy to expand the amount of data available for training. Analogously, several methods have been proposed to perform data augmentation in NLP, including Easy Data Augmentation (Wei and Zou, 2019), Back Translation (Sennrich et al., 2015), GPT-2 Augmentation (Anaby-Tavor et al., 2020), and `mixup` (Zhang et al., 2017). Kumar et al. (2020) applied some of these methods to pretrained transformer models and showed an average improvement in accuracy of 1-6%. However, the low-resource scenario was simulated by simply constraining the training data from large corpuses. It remains unclear how these methods might perform when used in realistic applications, where certain classes may be of very low frequency. One exemplary case concerns NLP analysis of online therapy sessions, where large amounts of patient-generated texts must be classified, but only well-trained specialists with relevant mental health domain knowledge can perform annotation manually to ensure clinical accuracy. In this study, we used a dataset from text message conversations between clients and therapists, previously used for detecting distorted thoughts (Tauscher et al., 2022). Besides the limitation in size, we found that some types of distorted thinking are very rare, resulting in worse classification performance. To address these issues, we investigate the extent to which data augmentation methods can improve performance of the best-performing BERT model from these experiments. We compare the utility of this augmentation approach to the use of a domain-specific pretrained language model, MentalBERT. In doing so, we evaluate the utility of data augmentation techniques and a domain-specific model to improve the identification of rare classes in the context of real-world data.

Our main contributions are as follows:

- We compared different augmentation methods in a low-resource dataset. We found improve-

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 68 - 75

ments with majority classes and that `mixup` can improve performance for rare classes.

- We adapted a domain-specific pretrained language model, MentalBERT, and showed the highest performance for majority classes, and better results for rare classes.
- We explored the hyperparameter α, controlling mixing proportions, for `mixup` and showed that a low α setting is helpful for dominant classes, and a high α for rare classes.

2 Low-resource Corpus

From our previous work (Tauscher et al., 2022), we utilized data from a randomized controlled trial of a community-based text-message intervention for individuals with serious mental illness (Ben-Zeev et al., 2020). Data were collected from 39 participants enrolled in the active intervention arm of this trial between December 2017 and October 2019. As part of the study, clients participating in standard care engaged with trained clinicians in text-message conversations up to three times a day for 12-weeks. In total, 14,312 messages were sent between clinicians and clients with 7,354 coming from clients. To build a predictive model for distorted thoughts, five common distortions were selected (Burns, 1999): Mental Filter (MF), Jumping to Conclusions (JC), Catastrophizing (C), Should Statements (SM), Overgeneralization (O). In addition, we added the label Any Distortion (AD), generated in accordance with the other assigned distortions. Two mental health specialists annotated all messages from clients by assigning these six categories, which are not mutually exclusive (Tauscher et al., 2022). This provided ground truth for labels. It is worth noting that any message could be identified as having multiple distortions, or no distortions at all, making this a multi-label multi-class problem. Table 1 shows the label frequency and inter-rater reliability.

	AD	**C**	**MF**	**JC**	**O**	**SM**
Frequency	24.4%	14.8%	8.6%	8.1%	3.6%	2.6%
kappa	0.51	0.44	0.33	0.53	0.46	0.39

Table 1: Label frequency and inter-rater reliability

3 Methods

Based on results by Tauscher et al. (2022), we used BERT as a starting point for our study, since it outperformed support vector machines and logistic regression (with L2 regularization), which had been used in prior work (Shickel et al., 2020; Shreevastava and Foltz, 2021). All models in this study were trained with the previously identified best hyperparameter settings for the dataset (Tauscher et al., 2022) (Section 3.1). Given the observed frequencies (Table 1), we combined results for six categories into three bins by frequency, to distinguish between effects on frequent and infrequent classes. The three bins are "high freq:AD,C", "medium freq:MF,JC", and "low freq:O,SM". For evaluation, we chose area under the precision-recall curve (AUPRC) over F_1 scores, because F_1 scores are special cases of AUPRC for a predefined cutoff and AUPRC is threshold-agnostic. For rare classes, the receiver operating characteristic curve (ROC) may lead to overly optimistic performance estimates, especially when class frequency drops to 1%, which is not the case with the precision-recall curve (Ozenne et al., 2015). Thus, we used AUPRC over others as our main metric. Macro-averaged AUPRC was calculated for each of the bins. This metric was also used to evaluate overall model performance.

We used two approaches to data augmentation, differing in the point at which augmentation occurs. The first involves directly augmenting the original text and outputting augmented examples as plain text, to be added to the original data (Section 3.2). The second approach involves augmentation in the hidden spaces of a deep neural network, and its outputs are vectors in the hidden space, rather than plain text (Section 3.3). For domain-specific model, we utilized a domain-specific pretrained language model with additional linguistic knowledge pertinent to the task at hand (Section 3.4).

3.1 BERT-based Classification

The baseline model we used is BERT (bert-base-uncased [1]) (Devlin et al., 2018). A classification layer was added on top of BERT's output and used for classifying all five cognitive distortions ("MF", "JC", "C", "SM", "O") and "AD". The maximum sequence length was set to 120 (word pieces).

The main framework for evaluation is 5-fold cross validation, and out-of-sample predictions were collected for the whole dataset. Following the original paper (Tauscher et al., 2022), we used the best hyperparameter settings for each of the iterations, as shown in Table 2. Also, losses were

[1] https://huggingface.co/bert-base-uncased

weighted inversely proportional to label frequencies.

Iteration	#1	#2	#3	#4	#5
number of epochs	14	14	10	14	8
dropout	0.2	0.3	0.1	0.2	0.2

Table 2: BERT hyperparameter settings

We repeated 5-fold cross validation five times with fixed folds but different random instantiations of the classification layer to assess the robustness of the results. This is the base setting for our experiments and was used across all other methods. This baseline model is labeled as "BERT (no aug)".

3.2 Augmentation of text data

3.2.1 EDA: Easy Data Augmentation

Wei and Zou (2019) proposed Easy Data Augmentation (EDA), which comprises of four main operations on the original text: Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). EDA was evaluated on five different tasks and showed an increased performance of 0.8% on average.

We adopted authors' recommended setting for the parameter α, 0.1, that controls the percentage of words in a sentence changed by each augmentation method. This is labeled as "BERT (EDA)".

3.2.2 Back Translation

Sennrich et al. (2015) proposed Back Translation for data augmentation, where sentences are first translated into another language and then back to the original language. This technique has been explored for the task of neural machine translation (Sugiyama and Yoshinaga, 2019). To generate new texts, we applied Back Translation with two intermediate languages: German and Spanish. During the augmentation, each original message was translated into German or Spanish and then back to English to get a corresponding message. Class labels of the original text were inherited. We did not repeat these experiments because we found little to no variation in generated texts upon repetition. The two backtranslation models are labeled as "BERT (BT:German)" and "BERT (BT:Spanish)".

3.2.3 GPT-2

Anaby-Tavor et al. (2020) propose using GPT-2 for data augmentation, by fine-tuning the model to generate text corresponding to a class of interest. Following their proposed approach, and using a publicly available GPT-2 model[2], we implemented two variations of GPT-2 for data augmentation.

Context-agnostic GPT-2: we first reconstructed our text messages as follows:

$$y_i[SEP]x_i[EOS]$$

for each of the messages i, where y_i indicates the label of a message, and x_i the message content. GPT-2 was then fine-tuned on this new structure of data for 20 epochs. New messages were generated by feeding in the prompt of "$y[SEP]$". This is labeled as "BERT (GPT-2: no context)".

Contextual GPT-2: Texts in our dataset are derived from conversations. To utilizing this contextual information, we reorganized inputs as follows:

$$y_i[SEP]x_{i-1}[SEP]x_i[EOS]$$

where x_{i-1} is the previous message. The GPT-2 model was then fine-tuned on this structure. Given the prompt of "$y_i[SEP]x_{i-1}[SEP]$", new messages were generated according to the class label y_i and and the preceding message for a representative example as context. This is labeled as "BERT (GPT-2: contextual)".

For text generation, we followed same steps described in Kumar et al. (2020). Due to computational time requirements, we did this once only.

3.3 Augmentation of Hidden Spaces: `mixup`

Zhang et al. (2017) proposed `mixup` for data augmentation. The authors claim that this method extends the training distribution by incorporating the prior knowledge that linear interpolations of feature vectors should lead to linear interpolations of the associated targets, providing data are modeled on vicinity relation across examples of different classes. `mixup` operates as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j,$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

where $\lambda \sim Beta(\alpha, \alpha)$ for $\alpha \in (0, +\infty)$. This paper did not examine the hyperparameter α across different NLP applications, with results reported only for Google speech commands, a dataset of 65,000 one-second utterances[3]. However, the authors did report improved results when using

[2] https://huggingface.co/gpt2

[3] https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

$\alpha = 0.3$ for this task, and in general proposed a small $\alpha \in [0.1, 0.4]$, based on results on ImageNet-2012. They also acknowledge that model error is less sensitive to large α when increasing model capacity. Sun et al. (2020) applied `mixup` to the transformer architecture and showed improvements on eight GLUE benchmarks. Across all of their experiments, α was fixed at 0.5, which is a reasonable extension from the originally proposed range (Zhang et al., 2017).

From the previous two studies (Zhang et al., 2017; Sun et al., 2020), it is not clear what hyperparameter setting of α should be used with other data sets. Given the probability density function controlled by α (demonstrated in Supplementary Figure 1), other settings when α is large may make more sense for scenarios in which we want to make two examples contribute more evenly. This leads to augmented examples lying in the margin between two categories, which may be appropriate for categories that are difficult to distinguish. In our case, the cognitive distortion dataset is relatively small compared with those evaluated previously, and some classes (O, SM) are quite rare. We wished to assess whether the `mixup` method could help with data augmentation in this context. We did an extended search in the hyperparameter space of α: 0.02, 0.2, 0.5, 1, 2, 4, and 8. The models are labeled as "BERT (`mixup`: $alpha = X$)".

3.4 Domain-Specific Model: MentalBERT

To investigate the utility of domain-specific models for transfer learning, we identified a domain-specific pretrained language model. Ji et al. (2021) describe MentalBERT and MentalRoBERTa, two language models developed specifically for mental health NLP. Starting with pretrained base models, and following standard BERT and RoBERTa pretraining protocols, MentalBERT and MentalRoBERTa were further pretrained on subreddits in the mental health domain, including "r/depression", "r/SuicideWatch", "r/Anxiety", "r/offmychest", "r/bipolar", "r/mentalillness", and "r/mentalhealth". These subreddits made up a pretraining corpus of over 13 million sentences. Upon evaluation, this additional pretraining improved performance in classifying mental conditions, including depression, stress, and anorexia. However, the evaluation sets used texts from online or SMS-like platforms, which were not fully annotated by specialists. In our work, we used MentalBERT, available from HuggingFace [4]. The same hyperparameters as the BERT model were used for comparison purposes. The baseline MentalBERT model is referred as "MentalBERT (no aug)". We also applied the best-performing data augmentation methods to MentalBERT, including back translation (Spanish) and explored some α settings for `mixup`.

4 Results

Performance for all models is shown in Table 3.

BERT: For the baseline BERT model, BERT (no aug), we obtain an AUPRC of 0.5179 for the most frequent classes (AD,C). When frequency decreases (classes MF,JC), the AUPRC also drops to 0.3718, and it drops further to 0.2139 for the rarest class of O,SM. This trend applies to all models. When data augmentation is applied to the base BERT model, we see improved results with different models. For the most frequent class of AD,C, back translation using Spanish achieves the highest AUPRC of 0.5208, followed by `mixup`: $\alpha = 0.02$. However, none of these results are significant improvements over baseline BERT. For the less frequent classes (MF,JC), back translation outperforms baseline BERT by 1.5%. `mixup` does not offer a performance boost here. When it comes to the rarest classes (O,SM), improvement is clearer: EDA, back translation (Spanish), and most settings of `mixup` can offer a boost in AUPRC. Among them, `mixup` ($\alpha = 4$) shows the biggest improvement in AUPRC by around 1.6%, which is statistically significant ($t(8) = 3.24, p\text{-}value = .012$ from t test). It is also notable that both GPT-2 based data augmentation methods decrease the performance of the base BERT model substantially (0.47 vs 0.52 for AD,C and 0.14 vs 0.21 for O,SM).

MentalBERT: When comparing MentalBERT results with BERT results, we can see improved performance for all classes, with the highest change for AD,C and MF,JC of 1.3%-1.8%. Similar to BERT models, performance is highly related to class frequencies, with highest being 0.5359 for the most frequent class of AD,C, dropping to 0.3846 for MF,JC then 0.2171 for O,SM. This trend holds for different augmentation settings. For augmentation effects, the base model performs best for both AD,C and MF,JC, as compared with augmented models. For rare class of O,SM, there is a small improvement from back translation (Spanish) of

[4] https://huggingface.co/mental/mental-bert-base-uncased

model	AUPRC (high freq:AD,C)	AUPRC (medium freq:MF,JC)	AUPRC (low freq:O,SM)	macro-AUPRC
BERT (no aug)	0.518 ± 0.0055	0.372 ± 0.0054	0.214 ± 0.0039	0.368 ± 0.0030
BERT (EDA)	0.517 ± 0.0062	0.378 ± 0.0071	0.228 ± 0.0091*	**0.374± 0.0067**
BERT (BT: German)	0.517	0.375	0.216	0.369
BERT (BT: Spanish)	**0.521**	**0.386**	0.222	0.376
BERT (GPT-2: contextual)	0.472	0.290	0.143	0.302
BERT (GPT-2: no context)	0.460	0.306	0.155	0.307
BERT (mixup: $\alpha = 0.02$)	0.519 ± 0.0013	0.372 ± 0.0026	0.218 ± 0.0078	0.370 ± 0.0041
BERT (mixup: $\alpha = 0.2$)	0.515 ± 0.0060	0.369 ± 0.0027	0.218 ± 0.0061	0.367 ± 0.0041
BERT (mixup: $\alpha = 0.5$)	0.510 ± 0.0058	0.367 ± 0.0058	0.213 ± 0.0034	0.363 ± 0.0033
BERT (mixup: $\alpha = 1$)	0.504 ± 0.0072	0.367 ± 0.0076	0.221 ± 0.0047	0.364 ± 0.0055
BERT (mixup: $\alpha = 2$)	0.505 ± 0.0043	0.366 ± 0.0046	0.222 ± 0.0054*	0.364 ± 0.0021
BERT (mixup: $\alpha = 4$)	0.505 ± 0.0048	0.367 ± 0.0027	**0.229 ± 0.0081***	0.367 ± 0.0038
BERT (mixup: $\alpha = 8$)	0.504 ± 0.0045	0.366 ± 0.0057	0.218 ± 0.0059	0.363 ± 0.0030
MentalBERT (no aug)	**0.536 ± 0.0029***	**0.385 ± 0.0059***	0.217 ± 0.0018	**0.379 ± 0.0032***
MentalBERT (BT: Spanish)	0.520	0.380	**0.222**	0.374
MentalBERT (mixup: $\alpha = 0.02$)	0.529 ± 0.0050*	0.379 ± 0.0031*	0.211 ± 0.0052	0.373 ± 0.0022*
MentalBERT (mixup: $\alpha = 0.2$)	0.523 ± 0.0033	0.382 ± 0.0049*	0.216 ± 0.0030	0.374 ± 0.0030*
MentalBERT (mixup: $\alpha = 1$)	0.520 ± 0.0064	0.381 ± 0.0056*	0.214 ± 0.0068	0.372 ± 0.0020*
MentalBERT (mixup: $\alpha = 4$)	0.515 ± 0.0028	0.379 ± 0.0021*	0.215 ± 0.0063	0.370 ± 0.0028
MentalBERT (mixup: $\alpha = 8$)	0.515 ± 0.0049	0.377 ± 0.0037	0.213 ± 0.0060	0.368 ± 0.0044

Table 3: AUPRC (mean ± std) for combined labels by frequency. *: significantly > BERT (no aug), unpaired t-test.

0.5%. None of the mixup configurations provide a benefit over the base MentalBERT model.

mixup: We explored an extensive range of the hyperparameter α with the BERT model. In Table 3, the best results usually come with a small α (0.02) for the dominant classes of AD,C and MF,JC. This best setting shows an increase of 1-2%. With an increasing α, the performance drops. For the rare classes of O,SM, a small α is no longer favored. The performance of AUPRC is not monotonic: with an increasing α, it first increases then drops, with its peak of 0.2285 at $\alpha = 4$. A similar trend is also observed for the MentalBERT model, although mixup did not perform best in this case.

Overall model performances is consistent with some of the preceding observations: (1) data augmentation improves overall performance, but only by a small margin; (2) in-domain pretraining of the language model (MentalBERT) provides the most improvement in performance; (3) for mixup, a small α is favored (0.02 for BERT and 0.2 for MentalBERT).

5 Discussion

We examined several data augmentation methods and explored their applications in BERT and MentalBERT for detecting distorted thinking in a modestly-sized set of text-based therapy messages. Grouping distortion classes by frequency, we found that most of data augmentation methods do not improve performance for frequent classes (frequency: 8-25%). For rare classes (3%), mixup significantly improved AUPRC results by 1.6%. In comparison, the domain-specific pretrained language model, MentalBERT, offered the highest benefit for dominant classes. However, MentalBERT also performs relatively poorly with rare classes. This may be due to the limited number of training examples. Another reason might be the fact that our text messages sometimes represent general conversations related to case management (e.g. appointment reminders) rather than the specific mental health related concerns that predominate in mental-health-related subreddits.

We also explored different settings for the hyperparameter α for the mixup method. For dominant classes, mixup favors a small α, which corresponds with previous work (Zhang et al., 2017). This indicates the model performs better with limited mixing of two random samples, generating cases where only one example predominates. In comparison, a larger α is favored for rare classes. According to Supplementary Figure 1, this means the model tends toward mixes in which the influence of individual texts is diluted, a possible way to create more variation in this low-resource scenario for the model to learn from. However, progressing to more extreme values ($\alpha = 8$) harms performance, and this cutoff point may change in other settings. Taken together, our results suggest that mixup is helpful for rare classes, but may compromise performance on frequent classes. Future work with mixup should include increasing the number of training epochs, since Zhang et al. (2017) sug-

label	Generated Text
JC	Yes you understand that it's incredibly frustrating and a lot of hard work but it's not at all stressful
C	Okay, i will do that, eventually

Table 4: GPT-2 generated text

gest that errors may be further reduced with more iterations of training.

Contrary to expectations, GPT-2-based data augmentation harmed performance in this context. It appears that GPT-2 generated texts (Table 4) do not express cognitive distortions as intended. This is likely because the data are not large enough to fully train a "distorted" GPT-2 model. Another reason may be that our prompts are not associated with distorted text by GPT-2. Designing better prompts may be a fruitful direction for future work.

6 Conclusions

We compared a range of data augmentation strategies and a domain-specific pretrained language model for their utility in improving identification of infrequently observed cognitive distortions. Using a domain-specific pretrained language model (MentalBERT) provided the greatest improvements, especially for dominant classes, whereas data augmentation did not improve performance with this model. In contrast, some data augmentation methods significantly improved performance with the base BERT model, but we did not find a method to improve performances for all classes universally, nor did we find a consistent hyperparameter setting to improve performance across these class frequencies. `mixup` appears helpful for rare classes, but a relatively large hyperparameter setting for α should be used. However, this may compromise the performance on frequent classes to some degree. Taken together our results suggest that the domain-specific model may be a better strategy for frequent classes, and that the best data augmentation strategy for infrequently observed classes varies across frequency ranges. As future work, two areas of interest include: (1) modified loss functions, such as the Label-Distribution-Aware Margin (LDAM) Loss (Cao et al., 2019) and Class-Balanced (CB) Loss (Cui et al., 2019), which have been proposed in the field of computer vision to address class imbalance; (2) unsupervised learning frameworks to address the inherent uncertainty of labels for augmented data, such as Confident Learning (Northcutt et al., 2021) and Unsupervised Data Augmentation (UDA) (Xie et al., 2020).

Acknowledgements

This work was supported by the UW Medicine Garvey Institute for Brain Health Solutions; National Institute of Mental Health grant (R56MH109554); and in part by the National Institutes of Health, National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (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

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

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. 1999. *Feeling Good: The New Mood Therapy*. Harper Collins, New York, NY.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

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*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

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 arXiv:2110.15621*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. 2015. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.

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

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S Yu, and Lifang He. 2020. Mixup-transformer: dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*.

Justin Tauscher, Kevin Lybarger, Xiruo Ding, Ayesha Chander, William Hudenko, Trevor Cohen, and Dror Ben-Zeev. 2022. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric Services (in review)*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

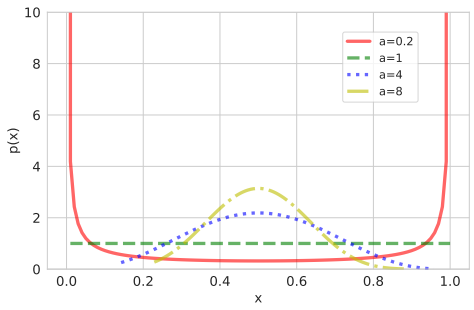
Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. *mixup*: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

A Appendix

Supplementary Figure 1

Figure 1: Probability Density Function of $Beta(\alpha, \alpha)$



In the paper of `mixup`, a special form of $Beta(\alpha, \beta)$ distribution was used where $\alpha = \beta$. The figure shows PDF of different α settings and this could affect the distributions of how the weights of two samples are assigned.