Inducing Positive Perspectives with Text Reframing

Caleb Ziems [*†] **Minzhi Li** [*◇] **Anthony Zhang** [†] **Diyi Yang** [†]
[†]Georgia Institute of Technology
{cziems, azhang305, dyang888}@gatech.edu
[◇]National University of Singapore
li.minzhi@u.nus.edu

Abstract

Sentiment transfer is one popular example of a text style transfer task, where the goal is to reverse the sentiment polarity of a text. With a sentiment reversal comes also a reversal in meaning. We introduce a different but related task called *positive reframing* in which we neutralize a negative point of view and generate a more positive perspective for the author without contradicting the original meaning. Our insistence on meaning preservation makes positive reframing a challenging and semantically rich task. To facilitate rapid progress, we introduce a large-scale benchmark, POSITIVE PSYCHOLOGY FRAMES, with 8,349 sentence pairs and 12,755 structured annotations to explain positive reframing in terms of six theoretically-motivated reframing strategies. Then we evaluate a set of state-of-the-art text style transfer models, and conclude by discussing key challenges and directions for future work. To download the data, see https://github.com/GT-SALT/positive-frames

1 Introduction

> Gratitude is not only the greatest of virtues, but the parent of all the others.
> — **Marcus Tullius Cicero**

Text style transfer (TST) has received much attention from the language technologies community (Hovy, 1987; Jin et al., 2020), where the goal is to change some attribute, like the sentiment of the text, without changing any attribute-independent content (Mir et al., 2019; Fu et al., 2018; Logeswaran et al., 2018). Some TST applications such as de-biasing (Pryzant et al., 2020; Ma et al., 2020) and paraphrasing (den Bercken et al., 2019; Xu et al., 2012) require meaning-preserving transformations, while political leaning (Prabhumoye et al., 2018), sentiment (Shen et al., 2017; Hu et al., 2017), and topical transfer (Huang et al., 2020) allow for a change in the underlying meaning. For instance, for a negative review, "*this was a bland dish*," we can use a sentiment TST model to create a more positive "*this was a tasty dish*," by swapping the word *bland* with *tasty*. Although the input's structure and attribute-independent content are preserved, the truth-conditional meaning is clearly altered.

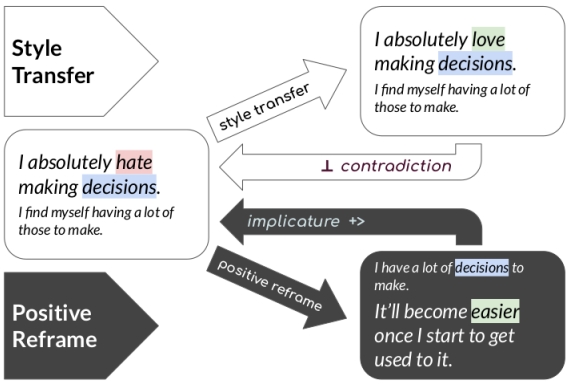

Figure 1: Positive reframing vs. negative-to-positive sentiment style transfer.

In this work, we introduce a closely related task—positive reframing—that differs from sentiment TST in important ways. We effectively *reframe* negative text by inducing a **complementary positive viewpoint** (e.g. *glass-half-full*), which nevertheless supports the underlying content of the original sentence. The reframe should implicate rather than contradict the source (see Figure 1), and the transformation should be motivated by theoretically justified strategies from from positive psychology (Harris et al. 2007; see Section 3).

To use the example from before, we could reframe "*this was a bland dish*" with the self-affirmation strategy and say "*I've made dishes that are much tastier than this one*." This reframed one still communicates the author's original intention by conversationally implicating that the dish was unsatisfying (Grice, 1975), but it shifts the focus away from the negative judgment and onto a positive and

[*]Equal contribution.

self-affirming perspective. Numerous studies have shown the positive effects of this and other reframing strategies on well-being and cognitive performance (Martens et al., 2006; Cohen et al., 2006; Good et al., 2003), which motivate this work.

Our main contribution is the design and implementation of a new *positive reframing* task. To facilitate research in this space, we introduce a parallel corpus of 8,349 reframed sentence pairs and 12,755 structured annotations for six theoretically-motivated re-write strategies. This is a significant contribution, especially since rich *parallel* corpora are scarce in TST tasks. Some related datasets exist for politeness (Madaan et al., 2020) and sentiment transfer (Shen et al., 2017; He and McAuley, 2016), but they lack this parallel structure. With only unaligned corpora, researchers are limited to unsupervised training paradigms, which notoriously fail to disentangle style from content, and thus also fail to preserve meaning (Lample et al., 2019). Using our parallel corpus, we examine how current state-of-the-art neural models work for positive reframing. We find that, supervised transformer-based neural models appear capable of rewriting a negative text without contradicting the original premise of that text. However, these models still struggle to generate reasonable positive perspectives, suggesting that our dataset will serve as a useful benchmark for understanding psychologically well-motivated strategies for augmenting text with positive perspectives.

2 Related Work

2.1 Style-Transfer

There is a longstanding interest in style transfer, starting with the early days schema-based systems (McDonald and Pustejovsky, 1985; Hovy, 1987), and then syntax-based (Zhu et al., 2010; Xu et al., 2016) and phrase-based machine translation (Xu et al., 2012; Wubben et al., 2012), into the age of end-to-end neural models. Recent works include supervised seq2seq tasks on parallel data (Rao and Tetreault, 2018; Fu et al., 2018) or pseudo-parallel data (Jin et al., 2019; Zhang et al., 2020b), as well as unsupervised generative modeling on non-parallel data (Hu et al., 2017; Shen et al., 2017), and semi-supervised techniques (Shang et al., 2019). Other ideas include domain adaptation (Li et al., 2019) or multi-task learning (Niu et al., 2018), zero-shot translation (Korotkova et al., 2019), unsupervised "delete and generate" approaches (Li et al., 2018; Sudhakar et al., 2019; Malmi et al., 2020; Madaan et al., 2020), and reinforcement learning (Zhang and Lapata, 2017; Wang et al., 2016).

Many existing datasets lack parallel structure, so the unsupervised setting is common in TST. Unfortunately, many of these methods still fail to disentangle style from content and adequately preserve the meaning of the original text (Lample et al., 2019). Autoencoders are particularly vulnerable to this shortcoming (Hu et al., 2017; Zhao et al., 2018), but some unsupervised machine translation techniques appear less vulnerable (Artetxe et al., 2018; Lample et al., 2018). In contrast, our positive reframing task requires source meaning-preservation and the introduction of *new* content and *new* perspectives, posing a unique challenge to unsupervised methods. We also provide a parallel corpus to train supervised models for this task.

2.2 Language and Positive Psychology

Positivity is contagious and can spread quickly across social networks (Coviello et al., 2014; Hatfield et al., 1993). Positive contagion in teams can reduce group conflict and improve group cooperation and even task performance (Barsade, 2002). Effective leaders also harness the power of positive reframing to promote company growth (Sy and Choi, 2013; Sy et al., 2005; Johnson, 2009; Masters, 1992) and beneficially shape negotiations (Filipowicz et al., 2011), customer relations (Dietz et al., 2004), decision making (Gächter et al., 2009; Druckman, 2001) and policy outcomes (Erisen et al., 2014). At an individual level, people who express optimism and gratitude are less likely to have depressive symptoms (Lambert et al., 2012) and more likely to experience emotional and psychological well-being (Carver et al., 1999; Watkins et al., 2008; Scheier et al., 2001).

On the other hand, fake expressions of positivity are correlated with negative brain activity (Ekman et al., 1990) and may actually be more harmful than helpful (Fredrickson, 2000; Fredrickson and Losada, 2005; Gross, 2013; Logel et al., 2009). That is why in our task it is essential that any *positively reframed* rephrased text remain true to the original premise of the source. In this way, our task is most similar to meaning-preserving transformations via parallel corpora from domains such as political argumentation (Chakrabarty et al., 2021), de-biasing (Pryzant et al., 2020; Ma et al., 2020), politeness (Madaan et al., 2020), and paraphrasing

(den Bercken et al., 2019; Xu et al., 2012).

3 Positive Reframing Framework

In this section, we present our psychologically-motivated taxonomy of positive reframing strategies. Instead of merely swapping antonyms for negative words or inserting unfounded positive language into a sentence, these strategies work to more fundamentally reconstruct the author's fixed, global, and ultimately harmful self-narratives, which are known in the literature as *cognitive distortions* (Burns, 1981; Abramson et al., 2002; Walton and Brady, 2020). Cognitive distortions include many exaggerated or irrational self-focused thoughts (Nalabandian and Ireland, 2019), such as dichotomous "all-or-nothing" thinking (Oshio, 2012), over-generalization (Muran and Motta, 1993), and catastrophizing (Sullivan et al., 2001). We can reconstruct these ideas using strategies from positive psychology (Harris et al., 2007). Each strategy is designed to promote a beneficial shift in perspective *without distorting the underlying context* of the author's situation.

Growth Mindset or, alternatively, the *incremental theory of personality* (Yeager et al., 2014; Burnette and Finkel, 2012), is the belief that one's skills and abilities are not immutable but can instead be changed and improved over time (Dweck, 2016); that one's willpower is an *abundant* rather than limited or exhaustible resource (Job et al., 2010, 2015); and that apparent setbacks like stress can be enhancing rather than debilitating (Crum et al., 2013). Instead of saying "*I'm such a lazy procrastinator*," a growth-mindset would say "*I'm determined to learn better time management*." This mindset has demonstrable benefits like improved performance on school tests (Good et al., 2003; Blackwell et al., 2007; Dweck and Yeager, 2019; Yeager et al., 2014).

Impermanence means understanding that negative experiences are finite and temporary, and that others have also experienced or even overcome similar forms of adversity. Someone might say "*since I failed this test, I must be too stupid for school*." An impermanence reframe could be "*This wasn't the test score I hoped for, but everyone slips up now and then.*" This category is also related to those proposed by Walton and Brady (2020): (1) focus on the "possibility of improvement," (2) recognize "specific, normal causes," and (3) understand "you're not the only one."

Neutralizing involves removing or rewriting negative phrases and terms so they are more neutral (Pryzant et al., 2020). Someone might complain that "*Wendy's customer service is terrible*." A neutralized reframe could be "*Wendy's customer service could use some improvement*."

Optimism does not mean to negate or deny the negative aspects of a situation, but instead to shift the emphasis to the more positive aspects of the situation, including expectations for a bright future (Carver et al., 2010). For example, if there is a negative emphasis, like in the sentence, "*I've completely worked myself to the bone this week, burning the candle at both ends... TGIF*," we can use optimism to shift the emphasis towards the positive as follows: "*It's been a long week, but now I can kick back, relax, and enjoy my favorite shows because it's the weekend*."

Self-affirmation means to assert a more holistic or expansive version of oneself by listing one's values, skills, and positive characteristics (Cohen and Sherman, 2014; Silverman et al., 2013). Positive psychology gives many examples like love, courage, hope, gratitude, patience, forgiveness, creativity, and humor (Harris et al., 2007). Reflecting on these values can bolster one's sense of integrity (see Self-Affirmation Theory; Steele 1988), can reduce depressive affect (Enright and Fitzgibbons, 2000), and can translate to increased performance on measurable tasks like exams (Martens et al., 2006; Cohen et al., 2006; Sherman et al., 2009).

Thankfulness can also be described more broadly as an "attitude of gratitude" (Emmons and Shelton, 2002). Adding more positive words that convey thankfulness or gratitude (e.g. appreciate, glad that, thankful for). For example, we can reframe the rhetorical question ,"*Is it sad that I don't wanna be at home and wish that work could call me in early*?" by expressing gratitude for career: "*I am thankful that I have a job that makes me want to get out of bed everyday*."

4 Data Collection

We sourced all of our data from the Twitter API, filtering tweets according to the hashtag `#stressed` due to a few reasons. Note that at the time of data collection and annotation, there were no publicly available datasets with annotated

Label Distribution	Count	Label	Description	ICC	Gen
25.4%	2,120	Growth Mindset	Viewing a challenging event as an opportunity for the author specifically to grow or improve themselves.	0.59	3.77
19.5%	1,625	Impermanence	Saying bad things don't last forever, will get better soon, and/or that others have experienced similar struggles.	0.60	4.03
36.1%	3,015	Neutralizing	Replacing a negative word with a neutral word. For example, "This was a terrible day" becomes "This was a long day."	0.32	3.53
48.7%	4,069	Optimism	Focusing on things about the situation itself, in that moment, that are good (not just forecasting a better future).	0.44	3.89
10.1%	841	Self-affirmation	Talking about what strengths the author already has, or the values they admire, like love, courage, perseverance, etc.	0.42	3.75
13.0%	1,085	Thankfulness	Expressing thankfulness or gratitude with key words like appreciate, glad that, thankful for, good thing, etc.	0.68	3.95

Table 1: Summary statistics for POSITIVE PSYCHOLOGY FRAMES. (*Left*) Distribution of the non-exclusive labels across all 8,349 annotations shows a preference for *optimism* and *neutralizing* strategies. (*Right*) The quality of annotations is shown by moderate Intra-class Correlation (ICC), with reasonable *genuineness* (Gen) metrics for 100 randomly sampled datapoints.

cognitive distortions, and the literature on distortion classification was still relatively unexplored (Simms et al., 2017; Shickel et al., 2020). We instead chose the simple keyword `#stressed` to signal the anxiety, negative affect, and hopelessness that has been shown to accompany cognitive distortions by prior work (Sears and Kraus, 2009).[1] Our decision to use Twitter was also motivated by the 280 character limit, which ensured that samples were short, focused expressions of relatively atomic ideas, as opposed to longer narrative-style texts from discussion platforms like Reddit's `r/rant`.

Our filtered collection of negative texts comes from a collection of over 1 million `#stressed` tweets written between 2012 and 2021, and it excludes any replies and retweets, any insubstantial tweets less than 30 characters, and any text containing a URL, which is often associated with spam (Zhang et al., 2012; Grier et al., 2010). After we removed other hashtags or Twitter handles from the text, we used TextBlob (Loria, 2018) to exclude any overtly positive texts with a non-negative sentiment score. Finally, to reduce any confounds between cognitive distortions and hate speech, and to make the human annotation task more agreeable for crowd-workers, we excluded examples that were flagged as offensive with over 80% confidence according to HateSonar (Davidson et al., 2017).

[1]We also considered *pet peeve*, *fml*, and other keywords but manual inspection revealed that these tweets were unlikely to contain cognitive distortions. In contrast, *stressed* hashtag provides a high precision data collection. We acknowledge this as a limitation and urge readers to keep this mind when interpreting our findings.

4.1 Annotation

We recruited crowdworkers to reframe 8,687 randomly-sampled texts with two workers assigned to each task, so we had two unique reframe annotations for every tweet. The annotators were encouraged to decide independently which reframing strategy to use, and they could combine multiple strategies in the same reframe. We simply asked annotators to record the strategies they selected. Additionally, they gave us, on a scale from 1-5, a score indicating how positive the original text was, and separately, how positive the text had become after they reframed it. Finally, we asked workers to mark advertisements, spam, or any text they felt they could not understand or effectively reframe. These examples were later removed from the corpus (see Appendix A for details).

In total, 204 workers participated in this task. Before they worked on the task, workers were asked to be familiar with our task by reading our provided reframing examples for each of the six strategies (Section 3), along with detailed annotation instructions. Then they had to pass a qualification test to show they can recognize different strategies in different reframing examples, with at least 5 out of 6 multiple-choice questions answered correctly.

We paid all annotators a fair wage above the federal minimum and both manually and programmatically inspected their work for quality (see Appendix A). After removing any poor-quality data, we were left with 8,349 reframed sentences. The strategy label distribution is given on the left side of Table 1, where a single reframe can have more than one strategy label.

4.2 Data Quality

To determine the reliability of the reframing strategy constructs, we randomly sampled 100 annotations from Section 4.1 and asked three annotators to consider both the original text and the reframed text, and then the annotators marked which of the six strategies were used in the given reframe. This allowed us to compute inter-annotator agreement scores for the strategy labels in Table 1. We observe the Intra-class Correlation for one-way random effects between the three raters and find moderate inter-rater agreement across these attribute categories (min 0.32; max 68). We also asked this second round of annotators to evaluate the *genuineness* of the reframes on a scale from 1-5. Our instructions explain that, with a more genuine reframe, it is more likely that someone in the original situation would say something similar. We find that, across all strategy labels, the average genuineness score is ~ 4 out of 5, so we know the data conforms reasonably well to our task instructions.

5 Positive Reframing

With POSITIVE PSYCHOLOGY FRAMES, we then examine how generative models work to automatically suggest a negatively-oriented self-narrative with a more positive shift in perspective without distorting any of the underlying meaning of that text. To do so will make use of encoder-decoder or conditional language models, as well as the six positive psychology strategies outlined in Section 3.

5.1 Task Formulation

Let $(s, t, \boldsymbol{\psi}_t)$ be a single annotation tuple in POSITIVE PSYCHOLOGY FRAMES for original source text s and positive reframe target t, which uses positive psychology strategies given by the multi-hot encoded vector $\boldsymbol{\psi}_t$. In the Positive Reframing task, our goal is to encode s and, at decoding time, produce t which makes use of $\boldsymbol{\psi}_t$ strategies and preserves the underlying meaning of s. Therefore, we formulate the problem as conditional generation and, during training, we maximize the standard language modeling objective

$$\frac{1}{N}\sum_{i=0}^{N}\log p(g_i|g_{0:i-1})$$

over the string

$$\begin{aligned}\boldsymbol{g} &= \{s, \boldsymbol{\psi}_t, t\} \\ &= \{\texttt{<BOS>}, s_1, s_2, ..., s_n, \\ &\quad \texttt{<STRG>}, \psi_{\texttt{grow}}, \psi_{\texttt{imp}}, ..., \psi_{\texttt{thank}}, \\ &\quad \texttt{<REFR>}, t_1, t_2, ..., t_m, \texttt{<EOS>}\}\end{aligned}$$

where g_i is the ith token in the string of length N, which contains the start token <BOS>, the tokenized source $s_{1:n}$, the tokenized reframe target $t_{1:m}$, and the binary tokens $\psi_{\texttt{grow}}, \psi_{\texttt{imp}}, ...$ indicating whether a particular strategy (e.g. growth mindset) was used in reframe t.

At decoding time, we consider three settings: *Unconstrained* generation $p(t|s)$, *Controlled* generation $p(t|s, \boldsymbol{\psi}_t)$, and a strategy *Prediction* form of generation $p(t, \boldsymbol{\psi}_t|s)$. Unlike in the *Unconstrained* setting, the *Controlled* generation is conditioned on the desired strategies $\boldsymbol{\psi}_t$. In the *Prediction* setting, the model will concurrently predict the strategies it used to generate its own reframe.

Note that, we introduce three different model settings here to capture how positive reframing assistance might be used by people in the real world. Specifically, the *Unconstrained* setting models reframing text directly without being aware of any specific strategy to use. The *Prediction* setting extends the unconstrained mode, i.e., produce the reframed text and also output the reframing strategies used in the reframing process spontaneously. The *Controlled* setting simulates the scenario of producing a reframed text with the help of concrete positive reframing strategies.

5.2 Experimental Setup

For ground truth training, development, and testing, we randomly partition the annotations using an 8:1:1 ratio, with 6,679 train, 835 development and 835 test data. We fine-tune the GPT and GPT-2 language models (Radford et al., 2019) as well as two Seq2Seq neural machine translation models – LSTM (Hochreiter and Schmidhuber, 1997) and CopyNMT (See et al., 2017) – and finally, two encoder-decoder models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). For all models, we use greedy decoding. As an ablation in the *Unconstrained* setting, we also test a *No-pretrain* condition for GPT-2 in which we randomly initialize the model parameters before fine-tuning.

Retrieval: We test two simple retrieval systems: *Random* retrieval of a reframed sentence from the training set, and SBERT (Reimers and Gurevych,

		Model	Automatic Evaluation R-1	R-2	R-L	BLEU	BScore	Δ TB	Avg. Len	Human Evaluation Meaning	Positivity	Fluency
Retrieval		Random	9.6	**3.6**	8.4	0.17	84.8	**0.36**	**20.0**	2.79	3.03	3.60
		SBERT	**15.2**	1.9	**12.8**	**1.47**	**87.6**	**0.36**	17.7	**3.45**	**3.97**	**4.16**
Few-shot		GPT-3	18.3	**3.4**	15.5	2.9	**88.2**	**0.44**	17.3	**3.73**	**4.17**	**4.27**
		GPT-Neo	**18.7**	**3.4**	**16.0**	**3.0**	**88.2**	0.40	**17.6**	3.69	4.16	4.21
Unconstrained	$p(t\|s)$	GPT	13.3	1.8	11.3	1.1	86.4	0.37	21.1	3.55	3.91	4.08
		GPT-2 No-pretrain	13.2	1.3	11.4	0.66	**89.6**	0.37	16.9	3.11	3.66	3.96
		GPT-2	20.9	4.6	17.7	4.2	88.5	0.35	20.0	3.58	4.01	**4.18**
		Seq2Seq-LSTM	15.7	1.4	12.4	0.73	85.6	**0.49**	25.8	3.33	4.15	4.10
		CopyNMT	20.8	5.0	18.0	4.0	85.7	0.32	16.1	3.57	3.69	3.91
		T5	27.4	9.8	23.8	8.7	88.7	0.38	**35.3**	4.09	3.79	4.06
		BART	**27.7**	**10.8**	**24.3**	**10.3**	89.3	0.23	24.4	**4.13**	**3.81**	4.15
Predict	$p(t, \psi_t\|s)$	T5	**27.5**	**10.5**	24.0	**11.0**	89.0	0.23	**25.1**	**4.10**	3.64	**4.11**
		BART	27.3	10.2	**24.1**	9.85	**89.4**	**0.32**	23.4	4.09	**3.95**	**4.11**
Control	$p(t\|s, \psi_t)$	T5	27.7	10.0	23.9	8.8	88.8	**0.36**	**35.0**	4.11	3.89	4.07
		BART	**28.8**	**10.9**	**25.1**	**10.1**	**89.6**	0.27	24.7	**4.23**	**4.07**	**4.27**
		Human	*100*	*100*	*100*	*100*	*100*	*0.35*	*17.4*	*3.80*	*3.82*	*4.18*

Table 2: **Positive reframing results** meausred by Meaning including ROUGE-1 (R-1), ROUGE-1 (R-2), ROUGE-L (R-L), BLEU, BERTScore (BScore), Positivity via Δ TextBlob (Δ TB) and Fluency. State-of-the-art models can generate meaning-preserving reframes in the unconstrained setting $p(t|s)$ and strategy-predictive setting $p(t, \psi_t|s)$ as well as when we condition the generation to use the reframing strategy from the ground truth $p(t|s, \psi_t)$. The best in-category performance is **bolded**; best overall performance is **highlighted**.

2019) retrieval, which finds the most similar t in train by cosine similarity and retrieves one of the corresponding ground-truth r from the training set.

Few-shot Learning: Brown et al. (2020) shows the few-shot capabilities of language models and especially larger models like GPT-3. We evaluate few-shot abilities of both GPT-3 and its open-source implementation, GPT-Neo (Black et al., 2021) using $k = 5$ exemplars (See Appendix C).

5.3 Evaluation

Following other style transfer work with a parallel corpus (Jhamtani et al., 2017; Xu et al., 2012), we evaluate our models for semantic similarity with the ground truth using the BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020a). Since there are two ground truth annotations per tweet, we take the maximum of the two scores and report the average across these maxima. We also report $\Delta TextBlob$ or the average change in sentiment score according to TextBlob (Loria, 2018). Finally, we conduct human evaluation in which 50 items are distributed to 3 raters who score the reframed sentences for three criteria, each on a scale from 1 to 5. The criteria include *Meaning Preservation* (Shang et al., 2019), our task-specific objective, as well as the *Positivity* and *Fluency* of the generated text, following the sentiment style transfer literature (Luo et al., 2019)

5.4 Results

Automatic Evaluation Across these metrics (Table 2, *left*) in the unconstrained generation setting, the BART model provided the highest quality of positive reframes, while GPT provided the worst quality with results similar to the *No-pretrain* version of GPT-2. The pre-trained version of GPT-2 was trained on English web text, while GPT was trained on works of fiction, so it appears that pre-training decisions can affect performance.

We tested the two best-performing models, T-5 and BART, on the controlled generation and strategy-prediction settings as well and found that the both models performed reasonably. Overall, controlled generation boosts performance, since the model can target the gold standard's strategies, but these improvements are only slight (see the *Controlled* part in Table 2). This warrants further

investigation: in Section 5.6, we explore models' ability to identify the underlying strategies given an existing reframe to understand whether models can make sense of these underlying constructs.

Unsurprisingly, all supervised models outperformed our simple retrieval baselines. Most interestingly, few-shot GPT-3 and GPT-Neo also **could not** match the supervised models in terms of overlap with the ground truth (ROUGE, BLEU, BERTScore), but they still achieved a comparable positive shift in sentiment (Δ TextBlob).

Human Evaluation Human judgments both support and elaborate on the automatic evaluation findings. For our best performing BART and T-5 models, the average scores are very high, even surpassing the quality of the *Human* gold standard in all of the unconstrained, predictive, and controlled settings. These systems most effectively induce a natural-sounding positive reframe while also *preserving the meaning* of the original text. This is critical: controlled BART model scored 4.07 in Positivity and 4.27 in Fluency while also achieving the winning Meaning preservation score.

In contrast with BART, the few-shot systems fail to preserve the meaning of the original sentence, despite their ability to articulately induce a more positive sentiment (Positivity scores up to 4.17; Fluency scores up to 4.27). Meaning preservation is absolutely critical for this task. From these results, we can conclude that, at the present time, supervised learning may be the most viable option for achieving reliable positive reframing results. POSITIVE PSYCHOLOGY FRAMES will facilitate ongoing efforts in this direction.

Qualitative Investigation Table 3 shows example reframes generated by our best controlled BART model, with one example for each strategy (for a similar comparison between *models*, see Table 5 in Appendix D). We see that, even without explicit lexical overlap between the generation and ground truth, the model reframes can still shift the cognitive distortions and negative outlook to a more positive perspective. In each of these examples, the model does so without losing the underlying meaning of the original text. Transformer-based models appear to be capable of solving our task with reasonable success. However, success can be highly variable (as evidenced by Table 5), so there is still room for significant improvement.

5.5 Error Analysis

We manually go through 100 randomly sampled model generations by our best controlled BART model, and summarize the main error classes here. We manually investigated 100 randomly sampled model generations by our best controlled BART model, and summarize the four largest error classes here. First, 26% of generations contained **(1) insubstantial changes**. These were especially prominent in the *neutralizing* strategy where the model would swap only a few negative words, like changing the phrase "*I hate it*" to "*I don't like it.*" On the other hand, some reframed generations were so drastically modified they contained **(2) contradictions to the premise** (9% of instances). For example, "*Feel like crying, this math class is impossible to pass*" was transformed into "*This math class is hard, but I know I can pass it*" – a failure of meaning preservation. More concerningly, the system can generate **(3) self-contradictions** (6%) like the phrase, "*I don't like opening up to people, but I'm glad I have the courage to do it.*" Finally, like many other NLG systems, our system can produce **(4) hallucinations** (2%) with unmotivated perspectives, like mentioning a *good night sleep* when the original post was about nosebleeds in the bath.

5.6 Frame Strategy Classification

In Section 5.4, we observed only slight performance gains when conditioning the generation based on the ground-truth reframing strategy (*Control* section in Table 2). For this reason, we take a closer look at whether models can reliably understand and classify the reframe strategies underlying a given source-reframe text pair. We formulate this problem as a multi-label multi-class classification task over sentence pairs (s, t). Given both the source text and positive reframe target in the annotation tuple (s, t) from POSITIVE PSYCHOLOGY FRAMES, we predict the multi-hot encoded strategy vector $\psi_t = [s_{\texttt{grow}}; s_{\texttt{imp}}; ...; s_{\texttt{thank}}]$ using transformer models. We experiment with a set of state-of-the-art classifiers, including BERT (Devlin et al., 2019), RoBERTA (Liu et al., 2019), and XLNet (Yang et al., 2019).

As shown in Table 4, all of the classification models can learn to recognize the thankfulness, optimism, and growth mindset strategies with moderate reliability ($F_1 > 0.60$). Although XLNet model cannot identify the neutralizing strategy very well, BERT and RoBERTa models can achieve an

Original Text	Strategy	Positive Reframe
(a) Could someone just give me like $1000? It would change my life, stressing about rent, bills and food money is just the worst.	growth	I need to learn how to manage my money better so that I don't have to stress about rent, bills and food money. *Time to focus on making a budget, so I don't have to stress about rent, bills, and food money.*
(b) I just went back to school today And I'm already stressed cause we have MIDTERMS NEXT WEEK !!! AND THIS WEEKEND – JAM PACKED :-(	impermanence	I just went back to school today and I'm already stressed because we have midterms next week and this weekend – JAM PACKED *Well, just think by the end of next week we'll be all done with the midterms!*
(c) talacre beach really need to sort there checking in routeen caravans are never ready 30mins later and still sat in the car waiting ..	neutralizing	The caravans at Talacre beach need to sort out their check-in process, so the kids don't get stuck in the car. *Talacre beach's check in process could do with some improvement to reduce the wait time for customers.*
(d) So glad that tomorrow is Friday. This has seriously been the longest week of my life! tired	optimism	I'm glad that tomorrow is Friday. This week has been long, but I'm looking forward to the weekend. *I'm glad the weekend is coming up, so I can rest.*
(e) Sometimes I get these impulses to just throw a tantrum. Like throw/break things, cry and scream.	self-affirmation	Sometimes I get these impulses to just throw a tantrum. Like throw/break things, cry and scream. But I'm strong, and I know I can handle it. *It's normal for to feel overwhelmed sometimes but I know I am strong to handle and go through it.*
(f) The fact that my mom's been calling me every night to calm me down from school says so much. I've needed it these past few days	thankfulness	I'm really lucky to have such a caring mum who is willing to call me every night when I'm having a hard time. *My mom has been calling me every night to calm me down from school. I've needed it these past few days. I'm thankful for her.*

Table 3: **Example reframes.** We examine one original text for each of the strategies in our dataset along with its BART-generated positive reframe (first row of each block) and its ground truth reframe (*italics* in second row of each block). Even when the generations differ from the ground truth, the model's reframes are largely successful at shifting the perspective while still maintaining the underlying meaning of the original text.

Strategy	BERT	RoBERTA	XLNet	Support
Thankfulness	0.71	0.69	0.71	109
Neutralizing	0.59	0.60	0.49	302
Optimism	0.72	0.71	0.72	400
Impermanence	0.55	0.55	0.54	157
Growth	0.61	0.63	0.67	221
Self Affirmation	0.43	0.44	0.39	76

Table 4: Strategy classification F1 scores

F1 score of around 0.6. The impermanence and self-affirmation strategies appear more challenging for all three models to identify. Overall, the results here show that this task is tractable: reframe strategies are **learnable** by various classification models. This further supports the reliability of our Positive Psychology framework, confirming what we found with human reliability metrics in Section 4.2. Although we mainly treat this frame strategy classification as a robustness check and deep dive into the role of framing strategies, this task can also be a novel NLP or computational social science application on its own, i.e., determining the positive reframing relation between a pair of sentences.

6 Discussion and Conclusion

This work introduces a new and challenging NLG task called *positive reframing*. The objective is to construct a more positive outlook as a way of rephrasing a negative source text such that the meaning of that source is preserved. Our parallel dataset, POSITIVE PSYCHOLOGY FRAMES, will serve as a benchmark that will enable sustained work on this task. We experiment with many of the leading style-transfer models and show that these models can learn to shift from a negative to a more positive perspective using a combination of strategies from positive psychology. Importantly, the best models are fluent and effective reframing systems that can learn to largely preserve the meaning of the original text, even under a perspective shift. However, these models still struggle to generate reasonable positive perspectives, and even the best models are still prone to errors. We discuss four key error classes: insubstantial changes, contradictions to the premise, self-contradictions, and hallucinations, as shown in Error Analyses in Section 5.5. Overall, this suggests that our dataset can serve as a useful benchmark for understanding well-motivated positive reframing strategies and equipping natural language generation systems with positive perspectives.

Future work can dive deeper into these issues by enforcing a stronger level of semantic equivalence between the generation and the source text

(Nie et al., 2019). Even with semantic equivalence constraints, it would be necessary to also allow for the injection of new positive perspectives. Methods ranging from guided sequence generation (Krause et al., 2020) or semantic attention-guided decoding (Nie et al., 2019) to pragmatic reconstruction (Shen et al., 2019) and persona consistency (Kim et al., 2020) may all be applicable in follow-up studies.

Acknowledgements

The authors would like to thank reviewers for their helpful insights and feedback. CZ is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655 and DY is supported by the Microsoft Research Faculty Fellowship. This work is funded in part by a grant from Amazon.

Ethics

Annotation. We followed the guidelines for ethical annotation practices and crowdsourcing that are outlined in (Sheehan, 2018), including paying workers a fair wage above the federal minimum. If workers contacted us with any questions or concerns, we responded promptly to them within 24 hours. In the task interface, in the header, we warned annotators that the content might be upsetting, and we gave the following recommendation: "*if any point you do not feel comfortable, please feel free to skip the HIT or take a break.*".

Deployment. Although this data is designed for pro-social outcomes (i.e. increasing positivity in text), there may be unexpected use-cases for this data, such as obfuscating impolite or even hateful data to avoid detection (ElSherief et al., 2021). The parallel structure of the data means it is also possible to invert the direction of the seq2seq task to introduce more negative or pessimistic perspectives into a positive source. This is not a particularly new risk, since sentiment style transfer can accomplish a similar outcome in this direction. Still, we will require interested parties to sign a data-use agreement that encourages only ethical uses of POSITIVE PSYCHOLOGY FRAMES.

References

Lyn Y Abramson, Lauren B Alloy, Benjamin L Hankin, Gerald J Haeffel, Donal G MacCoon, and Brandon E Gibb. 2002. Cognitive vulnerability-stress models of depression in a self-regulatory and psychobiological context.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly*, 47(4):644–675.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

Lisa S Blackwell, Kali H Trzesniewski, and Carol Sorich Dweck. 2007. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development*, 78(1):246–263.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jeni L Burnette and Eli J Finkel. 2012. Buffering against weight gain following dieting setbacks: An implicit theory intervention. *Journal of Experimental Social Psychology*, 48(3):721–725.

David D Burns. 1981. *Feeling good*. Signet Book.

Charles S Carver, Christina Pozo, Suzanne D Harris, Victoria Noriega, Michael F Scheier, David S Robinson, Alfred S Ketcham, Frederick L Moffat Jr, and Kimberley C Clark. 1999. How coping mediates the effect of optimism on distress: a study of women with early stage breast cancer.

Charles S Carver, Michael F Scheier, and Suzanne C Segerstrom. 2010. Optimism. *Clinical psychology review*, 30(7):879–889.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics.

Geoffrey L Cohen, Julio Garcia, Nancy Apfel, and Allison Master. 2006. Reducing the racial achievement gap: A social-psychological intervention. *science*, 313(5791):1307–1310.

Geoffrey L Cohen and David K Sherman. 2014. The psychology of change: Self-affirmation and social psychological intervention. *Annual review of psychology*, 65:333–371.

Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, and James H Fowler. 2014. Detecting emotional contagion in massive social networks. *PloS one*, 9(3):e90315.

Alia J Crum, Peter Salovey, and Shawn Achor. 2013. Rethinking stress: the role of mindsets in determining the stress response. *Journal of personality and social psychology*, 104(4):716.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *ArXiv preprint*, abs/1703.04009.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3286–3292. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joerg Dietz, S Douglas Pugh, and Jack W Wiley. 2004. Service climate effects on customer attitudes: An examination of boundary conditions. *Academy of management journal*, 47(1):81–92.

James N Druckman. 2001. Using credible advice to overcome framing effects. *Journal of Law, Economics, and Organization*, 17(1):62–82.

Carol Dweck. 2016. What having a "growth mindset" actually means. *Harvard Business Review*, 13:213–226.

Carol S Dweck and David S Yeager. 2019. Mindsets: A view from two eras. *Perspectives on Psychological science*, 14(3):481–496.

Paul Ekman, Richard J Davidson, and Wallace V Friesen. 1990. The duchenne smile: emotional expression and brain physiology: Ii. *Journal of personality and social psychology*, 58(2):342.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert A Emmons and Charles M Shelton. 2002. Gratitude and the science of positive psychology. *Handbook of positive psychology*, 18:459–471.

Robert D Enright and Richard P Fitzgibbons. 2000. *Helping clients forgive: An empirical guide for resolving anger and restoring hope.* American Psychological Association.

Cengiz Erisen, Milton Lodge, and Charles S Taber. 2014. Affective contagion in effortful political thinking. *Political Psychology*, 35(2):187–206.

Allan Filipowicz, Sigal Barsade, and Shimul Melwani. 2011. Understanding emotional transitions: the interpersonal consequences of changing emotions in negotiations. *Journal of personality and social psychology*, 101(3):541.

Barbara L Fredrickson. 2000. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, 14(4):577–606.

Barbara L Fredrickson and Marcial F Losada. 2005. Positive affect and the complex dynamics of human flourishing. *American psychologist*, 60(7):678.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Simon Gächter, Henrik Orzen, Elke Renner, and Chris Starmer. 2009. Are experimental economists prone to framing effects? a natural field experiment. *Journal of Economic Behavior & Organization*, 70(3):443–446.

Catherine Good, Joshua Aronson, and Michael Inzlicht. 2003. Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6):645–662.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37.

James J Gross. 2013. *Handbook of emotion regulation*. Guilford publications.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Alex HS Harris, Carl E Thoresen, and Shane J Lopez. 2007. Integrating positive psychology into counseling: Why and (when appropriate) how. *Journal of Counseling & Development*, 85(1):3–13.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11:689–719.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *ArXiv preprint*, abs/2011.00416.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.

V. Job, G. Walton, K. Bernecker, and C. Dweck. 2015. Implicit theories about willpower predict self-regulation and grades in everyday life. *Journal of personality and social psychology*, 108 4:637–47.

Veronika Job, Carol S Dweck, and Gregory M Walton. 2010. Ego depletionis it all in your head? implicit theories about willpower affect self-regulation. *Psychological science*, 21(11):1686–1693.

Stefanie K Johnson. 2009. Do you feel what i feel? mood contagion and leadership outcomes. *The Leadership Quarterly*, 20(5):814–827.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Elizaveta Korotkova, Agnes Luhtaru, Maksym Del, Krista Liin, Daiga Deksne, and Mark Fishel. 2019. Grammatical error correction and style transfer via zero-shot monolingual translation. *ArXiv preprint*, abs/1903.11283.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *ArXiv preprint*, abs/2009.06367.

Nathaniel M Lambert, Frank D Fincham, and Tyler F Stillman. 2012. Gratitude and depressive symptoms: The role of positive reframing and positive emotion. *Cognition & emotion*, 26(4):615–633.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations,*

ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Christine Logel, Emma C Iserman, Paul G Davies, Diane M Quinn, and Steven J Spencer. 2009. The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*, 45(2):299–312.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5108–5118.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2:269.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Andy Martens, Michael Johns, Jeff Greenberg, and Jeff Schimel. 2006. Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, 42(2):236–243.

Mark A Masters. 1992. The use of positive reframing in the context of supervision. *Journal of Counseling & Development*, 70(3):387–390.

David D. McDonald and James D. Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, Switzerland. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Elizabeth M Muran and Robert W Motta. 1993. Cognitive distortions and irrational beliefs in post-traumatic stress, anxiety, and depressive disorders. *Journal of Clinical Psychology*, 49(2):166–176.

Taleen Nalabandian and Molly Ireland. 2019. Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or Friends. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 62–73, Minneapolis, Minnesota. Association for Computational Linguistics.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Atsushi Oshio. 2012. An all-or-nothing thinking turns into darkness: Relations between dichotomous thinking and personality disorders 1. *Japanese Psychological Research*, 54(4):424–429.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael F Scheier, Charles S Carver, and Michael W Bridges. 2001. Optimism, pessimism, and psychological well-being.

Sharon Sears and Sue Kraus. 2009. I think therefore i om: Cognitive distortions and coping style as mediators for the effects of mindfulness meditation on anxiety, positive and negative affect, and hope. *Journal of clinical psychology*, 65(6):561–573.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4937–4946, Hong Kong, China. Association for Computational Linguistics.

Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

David K Sherman, Geoffrey L Cohen, Leif D Nelson, A David Nussbaum, Debra P Bunyan, and Julio Garcia. 2009. Affirmed yet unaware: exploring the role of awareness in the process of self-affirmation. *Journal of personality and social psychology*, 97(5):745.

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.

Arielle Silverman, Christine Logel, and Geoffrey L Cohen. 2013. Self-affirmation as a deliberate coping strategy: The moderating role of choice. *Journal of Experimental Social Psychology*, 49(1):93–98.

Taetem Simms, Clayton Ramstedt, Megan Rich, Michael Richards, T Martinez, and C Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 508–512. IEEE.

Claude M Steele. 1988. The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in experimental social psychology*, volume 21, pages 261–302. Elsevier.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Michael JL Sullivan, Wendy M Rodgers, and Irving Kirsch. 2001. Catastrophizing, depression and expectancies for pain and emotional distress. *Pain*, 91(1-2):147–154.

Thomas Sy and Jin Nam Choi. 2013. Contagious leaders and followers: Exploring multi-stage mood contagion in a leader activation and member propagation (lamp) model. *Organizational Behavior and Human Decision Processes*, 122(2):127–140.

Thomas Sy, Stéphane Côté, and Richard Saavedra. 2005. The contagious leader: impact of the leader's mood on the mood of group members, group affective tone, and group processes. *Journal of applied psychology*, 90(2):295.

Gregory M Walton and Shannon T Brady. 2020. "bad" things reconsidered. In *Applications of social psychology*, pages 58–81. Routledge.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 4270–4271. AAAI Press.

Philip C Watkins, Lilia Cruz, Heather Holben, and Russell L Kolts. 2008. Taking care of business? grateful processing of unpleasant memories. *The Journal of Positive Psychology*, 3(2):87–99.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

David Scott Yeager, Rebecca Johnson, Brian James Spitzer, Kali H Trzesniewski, Joseph Powers, and Carol S Dweck. 2014. The far-reaching effects of believing people can change: implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of personality and social psychology*, 106(6):867.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xianchao Zhang, Shaoping Zhu, and Wenxin Liang. 2012. Detecting spam and promoting campaigns in the twitter social network. In *2012 IEEE 12th international conference on data mining*, pages 1194–1199. IEEE.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Data Quality-Control Methods

We used programmatic methods to ensure high-quality reframing annotations at submission time. Workers could not submit their task if the reframe: (1) contained fewer than 3 word types; (2) had a length less than 25% of the original text; (3) had more than 3 repetitions of a single bigram; or (4) was too similar to the original text, with a token Jaccard Similarity greater than 90%. Furthermore, we used the LanguageTool API[2] to prompt workers to fix any grammatical mistakes in their writing. Cumulatively, these heuristics greatly improved the annotation quality. Later, in the post-processing stage, we employed additional programmatic measures as well as manual quality-checks to filter out the unsatisfactory examples. This process was iterated after each batch, with a batch size of 100. First, one of the authors manually checked any sentences where annotators had scored the *original text* with a postivity score greater than 3 (out of 5). If that author found that the text was not negative enough or did not contain the requisite *cognitive distortions* to warrant a substantial reframing, the sentence was removed from the corpus. Next, we considered all *neutralizing* reframes with a score less than 4 (out of 5). If the text was not effectively neutralized, we removed the sentence from the corpus. Then we considered all annotations containing the first person pronoun *you*. If the text abandoned the author's first-person voice and shifted into a 3rd-person critique or commentary (e.g. "*I feel hopeless*" → "*you should find hope*"), then we removed this from the corpus. Finally, we grouped the annotations by Worker ID and, for each worker, scanned the top 10 annotations. If the annotator produced poor quality work, we removed the examples and blocked the worker from future tasks. After a last pass through the data to manually correct noticeable punctuation and grammar errors, we were left with our cleaned corpus of 8,349 reframed sentences.

B Task Interface

Figure 2 shows the Instructions we gave to the Amazon Mechanical Turk (MTurk) workers. Figure 3 shows the examples we displayed for each reframe strategy. Figure 4 shows the MTurk HIT interface that we used for the Section 4.1 task to collect positive reframes with their associated strategies as well as the positivity scores for both the original TEXT and the REFRAME. Figure 5 shows the interface for the Section 4.2 task where we collected new strategy labels for prior annotations to compute inter-annotator agreement scores.

[2] api.languagetoolplus.com/v2/check

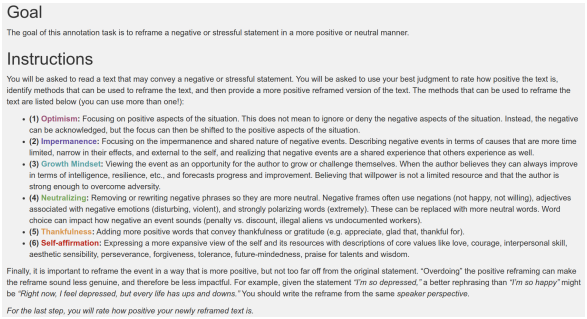

Figure 2: Instructions for the Positive Reframing HIT.

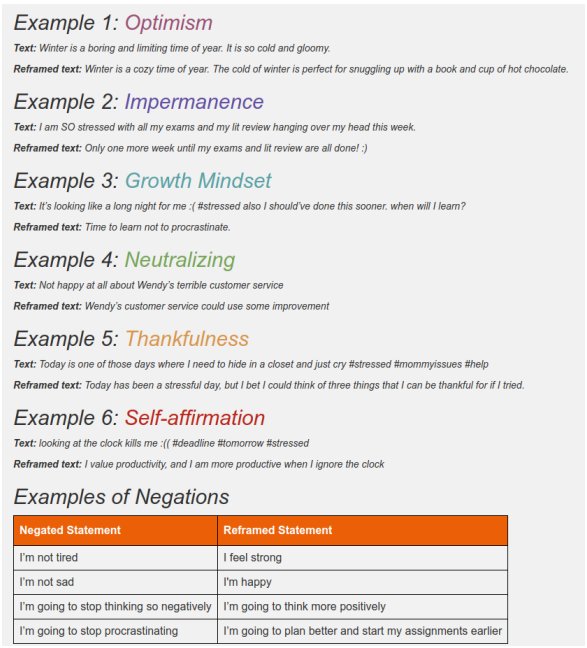

Figure 3: Example reframes.

C Few-shot Learning Setting

Following (Han et al., 2018; Baldini Soares et al., 2019) and others, we consider 5-shot learning. We pull 5 representative exemplars from the training set to indicate a range of strategies:

NEGATIVE: "I have a huge project due tomorrow morning. But where do I have to be, a stupid basketball game dumb"

POSITIVE: "I should plan ahead next time so that my basketball game does not conflict too closely with my projects."

NEGATIVE: "This has been like the worst week ever im so done with everything. sick tired"

POSITIVE: "I made it to the end of the most challenging week ever!"

. . . . . .

NEGATIVE: "Ugh my mac is starting to slow up and I need to figure out how to defragment the hard drive..."

POSITIVE: "I need to defragment the hard drive to speed up my mac. Good thing I'm smart, and I know I can do this."

. . . . . .

NEGATIVE: "I am SO stressed with all my exams and my lit review hanging over my head this week."

POSITIVE: "Only one more week until my exams and lit review are all done!"

. . . . . .

NEGATIVE: "I am the only person I know who writes a healthy grocery list and plans meals when I am stressed:(CantSleep"

POSITIVE: "I'm so thankful that I am still able to eat healthy even when I'm stressed."

. . . . . .

D Example Reframes

In Table 5, we compare examples of model-generated reframes from different models. The examples are structurally and semantically diverse, which may suggest that different architectures could serve as complementary systems in a broader effort to introduce a range of positive perspectives in text. However, the generations are not perfect. In this particular example, CopyNMT and GPT-2 fail to integrate the key concept of the "heavy workload" into their reframe generations. There is still significant room to improve upon these models in future work.

We were also interested in the success of the *Controlled* generations $p(t, \psi_t|s)$. Do models appropriately follow a given strategy when the generation is conditioned on that strategy? Table 6 shows different reconstructions of the same text using different controlling strategies. Qualitatively, we can see that controlled generation is a reasonable method for influencing the most prominent strategy. The 'thankfulness', 'growth' and 'self-affirmation' strategies are particularly distinct.

Original Text		Always stressing and thinking about loads of things at once need I take it one at a time overload stressed need to rant.
Ground Truth		Loads of things on my mind, I need to make a list, prioritise and work through it all calmly and I will feel much better.
Model Generation	**Random**	Annoyed because the USPS lost 1 of the 8 flat rate boxes of books I shipped from California to Kentucky. Hope it shows up :-/
	SBERT	Really stressed at the moment but I know that this will pass and I'll get everything sorted out.
	Seq2seq-LSTM	i have a lot of things to do it is a good night's sleep and i can get through it.
	CopyNMT	Always stressing and thinking easy of things of things at things at all, but I know that I need it.
	GPT	i ' m so relieved that i ' m going to get to organize my thoughts in such a short time ! i hope i am done with the list of things i have to do today , so that i can go back to bed .
	GPT-2 No-pretrain	I'm going to try to seek a new job to get some newspect me through today.
	GPT-2	I should be more energetic in my thoughts, and not worry too much.
	T5	I am stressed and thinking about loads of things at once, I need to take it one at a time stressed need to rant. I'm sure I'll be able to focus on it.
	BART	I am stressed and thinking about loads of things at once, but I am going to take it one at a time.

Table 5: **A model comparison** for reframing the same text. Here, we observe significant diversity in the generations, suggesting that these different architectures could serve as complementary systems in an effort to restruture and transform negative text on the web.

Original Text		I'm doing really well in some class and bad in other classes I hate this.
GPT-2 Generations	**Growth Mindset**	I'm doing really well in some class and bad in other classes I hate this. But I'm doing better than I was at the age of 11.
	Impermanence	I'm having a hard time in some classes and I hope to get through the rest of the semester.
	Neutralizing	I'm doing really well in some class and bad in other classes I hate this. But I'm going to go clear it out.
	Optimism	I need to take a good hard look at myself and realize that I am quite capable of doing better in other classes.
	Self-affirmation	I'm really happy about my class, I'm doing really well in some class, bad or not, I can do better in another.
	Thankfulness	I'm so lucky that I'm able to get my classes to a good start and get off the couch.
T5 Generations	**Growth Mindset**	I'm doing really well in some class and bad in other classes I hate this, but I'll be able to do it again.
	Impermanence	I'm doing really well in some class and bad in other classes I hate this. But I'll be able to do my best in the future.
	Neutralizing	I'm doing really well in some class and bad in other classes I hate this. But I can't wait to see if I can do it.
	Optimism	I'm doing really well in some class and bad in other classes I hate this. I hope I'll be able to do better in the future.
	Self-affirmation	I'm doing really well in some class and bad in other classes I hate this. It's a good thing to do, but I'll be able to do it.
	Thankfulness	I'm doing really well in some class and bad in other classes I hate this. But I can't wait to see if I can do it.
BART Generations	**Growth Mindset**	I'm doing really well in some class and bad in other classes. I'm going to try to improve my grades.
	Impermanence	I'm doing really well in some class and bad in other classes, but I'm sure it will all work out in the end.
	Neutralizing	I'm doing really well in some class and bad in other classes I don't like this.
	Optimism	I'm doing really well in some class and bad in other classes, but I'm sure it will all work out.
	Self-affirmation	I'm doing really well in some class and bad in other classes, but I know I can do better.
	Thankfulness	I'm doing really well in some class and bad in other classes, but I'm thankful that I have the opportunity to study.

Table 6: **A model comparison** for reframing the same text using different controlling strategy. Here, we observe models can learn some information from the input strategy label and make distinctive generations, especially for the 'thankfulness', 'growth' and 'self-affirmation' strategies.

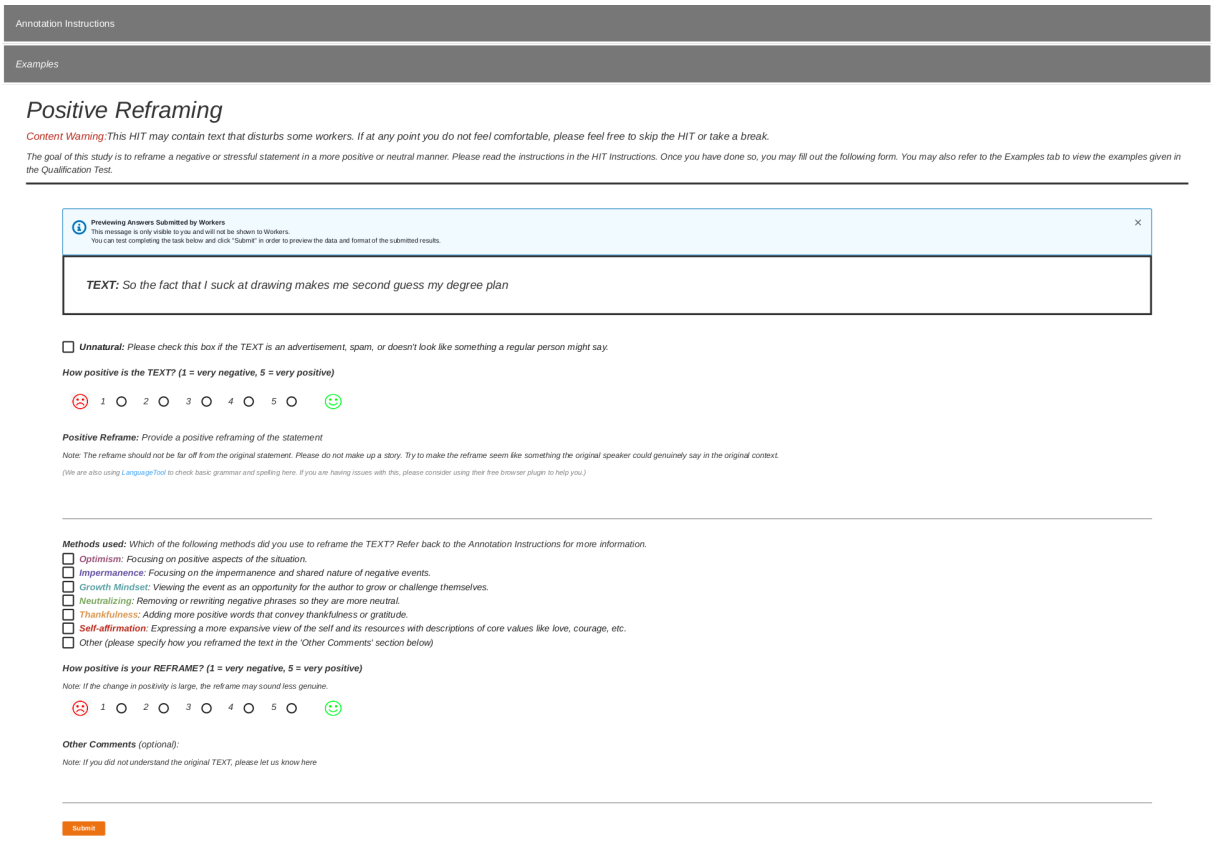

Figure 4: Amazon Mechanical Turk interface used to collect positive reframes (in Section 4.1).

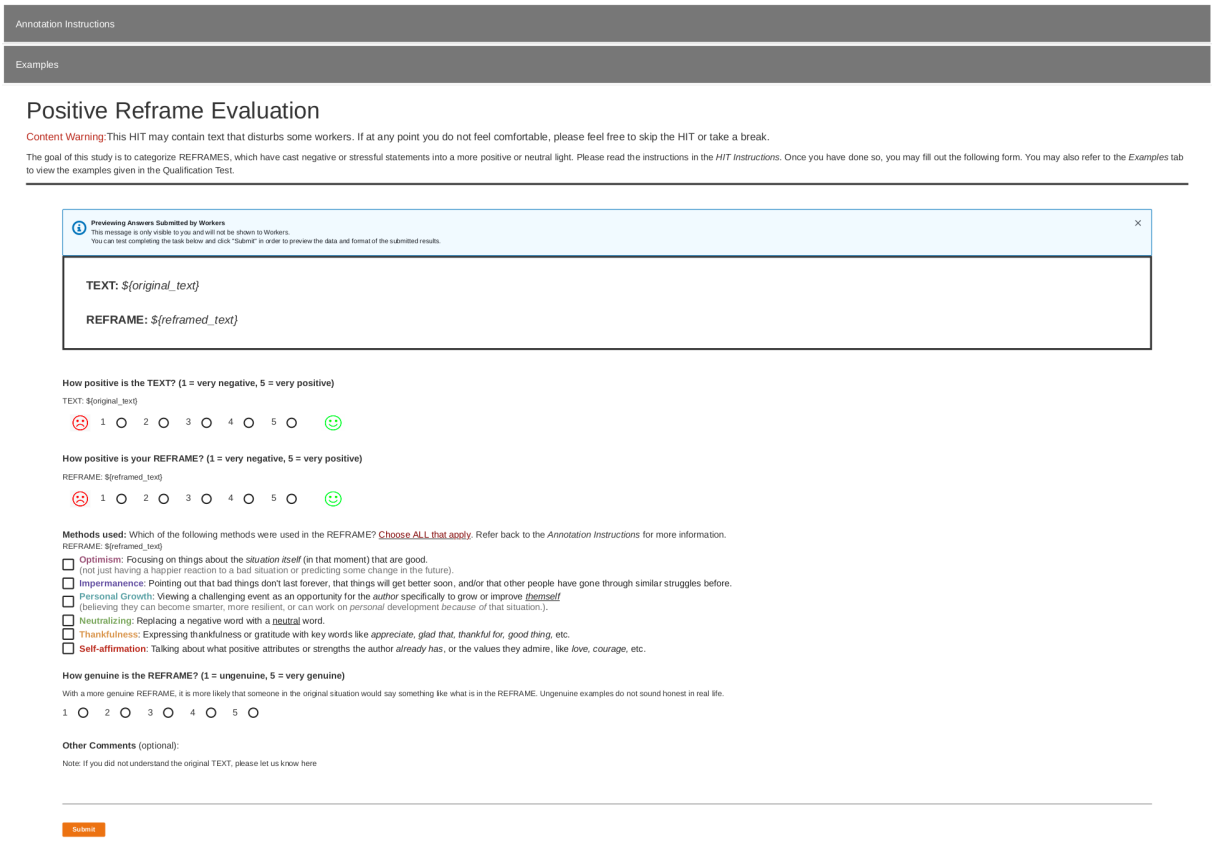

Figure 5: Amazon Mechanical Turk interface used to find inter-annotator agreement for the taxonomy (in Section 4.2).