APPDIA: A Discourse-aware Transformer-based Style Transfer Model for Offensive Social Media Conversations

Katherine Atwell[*], Sabit Hassan[*], Malihe Alikhani
Computer Science Department, School of Computing and Information
University of Pittsburgh, Pittsburgh, PA
{kaa139,sah259,malihe}@pitt.edu

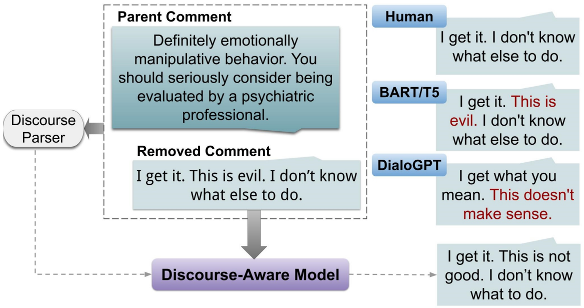
Abstract

Using style-transfer models to reduce offensiveness of social media comments can help foster a more inclusive environment. However, there are no sizable datasets that contain offensive texts and their inoffensive counterparts, and fine-tuning pretrained models with limited labeled data can lead to the loss of original meaning in the style-transferred text. To address this issue, we provide two major contributions. First, we release the first publicly-available, parallel corpus of offensive Reddit comments and their style-transferred counterparts annotated by expert sociolinguists. Then, we introduce the first *discourse-aware* style-transfer models that can effectively reduce offensiveness in Reddit text while preserving the meaning of the original text. These models are the first to examine inferential links between the comment and the text it is replying to when transferring the style of offensive Reddit text. We propose two different methods of integrating discourse relations with pretrained transformer models and evaluate them on our dataset of offensive comments from Reddit and their inoffensive counterparts. Improvements over the baseline with respect to both automatic metrics and human evaluation indicate that our discourse-aware models are better at preserving meaning in style-transferred text when compared to the state-of-the-art discourse-agnostic models.

1 Introduction

Disclaimer: Due to the nature of this work, figures and examples may contain offensive phrases.

The spread of offensive and hateful content on social media can be detrimental to users' psychological well-being (Waldron, 2012; Gülaçtı, 2010). Anonymity on platforms such as Reddit can further embolden users to post such hateful content (Ascher, 2019). Further, the sheer volume of content on popular social media platforms can render the human moderation process ineffective (Hassan et al., 2022) or psychologically damaging for moderators (Dosono and Semaan, 2019) and calls for AI systems that can mitigate this problem.



Definitely emotionally manipulative behavior. You should seriously consider being evaluated by a psychiatric professional.

I get it. This is evil. I don't know what else to do.

I get it. I don't know what else to do.

I get it. This is evil. I don't know what else to do.

I get what you mean. This doesn't make sense.

I get it. This is not good. I don't know what to do.

Figure 1: Example of instances where pretrained language models either fail to remove offensiveness (BART/T5) or drastically alter the intended meaning (DialoGPT) when fine-tuned on our style transfer task

AI moderation of social media by simply removing content classified as offensive (Zampieri et al., 2020; Hassan et al., 2021) may reduce diversity in online conversations and deter users from using the platform (Jhaver et al., 2019). Our exploration reveals that many comments removed by moderators on Reddit contain contributions to the discourse beyond their offensive content. Rather than simply removing these comments from social media platforms, they can be turned into inoffensive statements by using alternative words, removing profanity, or paraphrasing certain parts, while preserving the overall semantic content.

We approach the problem of eliminating offensiveness from text while preserving original semantic content as a supervised *style-transfer* task, where offensive text is transferred to inoffensive text. As a first step towards this goal, we create the first publicly-available, expert-annotated style transfer corpus for Reddit data, which contains offensive comments that include certain lexical items and more subtle instances that are implicit and grounded in context. This differentiates our work from unsupervised approaches are mostly

Proceedings of the 29th International Conference on Computational Linguistics, pages 6063–6074
October 12–17, 2022.

good at handling instances with explicit lexical cues (Nogueira dos Santos et al., 2018).

Although large pretrained transformer models have been successfully deployed for generation tasks, these models come with the risk of either failing to generate desired output or obfuscating the source passage's meaning while still producing coherent text (Bender et al., 2021). In our work, we target the issue of content preservation using *discourse frameworks*, which have been successfully employed for various generation tasks (Maskharashvili et al., 2021; Xu et al., 2022; Bosselut et al., 2018), but have not been employed in style transfer models. We hypothesize that integrating discourse coherence frameworks within transformer-based style transfer models can contribute to better preservation of semantic content, specifically for short social media comments.

We study our hypothesis with a small pilot annotation of style-transferred text produced by pretrained transformer models. Figure 1 shows examples of the issues described above in our style transfer task, where BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) do not remove offensiveness from the original comment, but DialoGPT (Zhang et al., 2020b) significantly alters the original semantic content. We observe that coherence relations between a comment and its reply are not preserved under style transfer in cases where offensiveness is removed. For instance, the removed comment refers to *"emotionally manipulative behavior"* in the parent comment with *"This is evil"*, exhibiting the behavior of *"Same-Unit"* discourse relation, which is not preserved in the style-transferred text generated by DialoGPT.

To test our hypothesis, we provide the following contributions:

- Collect a dataset[1] of ~2K offensive comments from Reddit that are annotated by expert sociolinguists with inoffensive counterparts. Our data also contains parent comments/posts and are tagged with discourse relations, making it the first publicly available dataset of its kind.

- Propose two approaches for integrating discourse relation frameworks with pretrained transformer models: i) using Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008; Webber et al., 2019) relations within a single comment, and ii) parsing a comment and the text it is responding to using the Rhetorical Structure Theory Discourse Treebank (Mann and Thompson, 1988).

The results for both discourse-aware approaches indicate improvement in content preservation over the pretrained baselines, providing support for our hypothesis and for the use of discourse frameworks to preserve meaning in style-transfer tasks.

2 Related Work

Paraphrase generation is a well-studied problem that has yielded large datasets such as the PDTB paraphrase database (Ganitkevitch et al., 2013), WikiAnswer (Fader et al., 2013), ParaNMT (Wieting and Gimpel, 2018), and the MSCOCO dataset (Lin et al., 2014a). Recent works in the related but relatively new field of style transfer primarily target sentiment transfer (Li et al., 2018b; Yu et al., 2021), formality transfer (Chawla and Yang, 2020) or expertise transfer (Cao et al., 2020). Very few works have targeted transferring style from offensive to inoffensive text, with Nogueira dos Santos et al. (2018) and Cheng et al. (2020) being notable exceptions. Our dataset differs from the aforementioned works in multiple ways. Ours is the first publicly available dataset that contains offensive Reddit comments that are rewritten by experts, paired with parent comment/post, and automatically tagged with discourse relations. Further, both Nogueira dos Santos et al. (2018) and Cheng et al. (2020) derive their datasets from political subreddits (Serban et al., 2017), while our data encompasses subreddits on personal views, question-answer discussions, and gender rights in addition to political subreddits.

Development of pretrained language models (PLM) such as BART (Lewis et al., 2020) has changed the landscape of natural language generation research and we are witnessing a shift toward controllable text generation (Zhang et al., 2022; Zeldes et al., 2020; Ribeiro et al., 2021; Ziegler et al., 2019). Discourse relations have been proposed as a possible controlled generation method, shown to aid extractive and abstractive summarization (Cohan et al., 2018; Xu et al., 2020), text generation from meaning representations (Maskharashvili et al., 2021), and question answering with logical reasoning or complex answers (Huang et al., 2021; Xu et al., 2022). Discourse-aware models have also been shown to generate more coherent

[1] https://github.com/sabithsn/APPDIA-Discourse-Style-Transfer

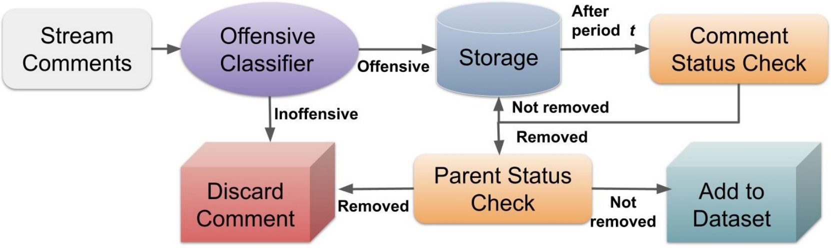


Figure 2: Our data collection pipeline for obtaining removed comments from Reddit that are offensive.

texts (Bosselut et al., 2018) within a reinforcement learning setting. Our work integrates both RST-DT and PDTB frameworks with pretrained transformer models and provides a comparison of the relative efficacy of the two frameworks for a generation task. Within the context of style transfer, recent works have focused on classification and reconstruction loss (Nogueira dos Santos et al., 2018; Chawla and Yang, 2020) in semi-supervised/unsupervised setting, use of copy mechanism (Jhamtani et al., 2017), and coherence classifier (Cheng et al., 2020) to guide the style transferred text. To our knowledge, our work is the first to utilize discourse coherence frameworks for style transfer.

3 Data Collection and Annotation

In order to reduce offensiveness in text, we create an expert-annotated dataset of offensive comments and their style-transferred counterparts. In this section, we first describe our pipeline for collecting and curating a set of offensive comments from Reddit. Then, we describe our annotation process for reducing offensiveness in these comments.

3.1 Data Collection Pipeline

First, we stream 14 subreddits spanning topics of politics, personal views, question-answer discussions, and gender rights for new comments using PRAW[2]. The streamed comments are then tagged for offensiveness using a BERT model (Devlin et al., 2019) fine-tuned on the OLID dataset (Zampieri et al., 2019a), which consists of 14K tweets annotated for offensiveness and was used for the SemEval 2019 (Zampieri et al., 2019b) shared task. If a comment is tagged as inoffensive by the classifier, we remove it from our data. As our initial exploration revealed that a large portion of the removed comments on Reddit (> 60%) may not be offensive and may have been removed due to violation of subreddit-specific rules, the exclusion of inoffensive comments from the data is essential for a feasible annotation process. Our manual annotation, as described in the next section, eliminates any false positive bias that the classifier may have. More details about the classifier can be found in the Appendix.

After running the classifier, the body and metadata of comments that are tagged as offensive, are stored locally. We then periodically check the accessibility status of these comments on Reddit. If it has been removed by moderators, we query Reddit for the *parent* comment or post that it is in response to. If the comment is a reply to another comment, then the comment it is replying to is considered the parent, and if the comment is a top-level comment, i.e, a direct reply to a post, then the post is considered the parent. If the parent has been deleted or removed, the comment is discarded. Otherwise, we add the comment, along with the parent, to our dataset. Our data collection pipeline is summarized in Figure 2. After filtering out very long comments, we end up with a pool of 5.8K comments for annotation.

3.2 Data Annotation

The 5.8K comments obtained from our data collection pipeline are annotated by three expert sociolinguists. The annotators are paid 30 USD per hour and the Institutional Review Board (IRB) protocol, as defined by our institution, the University of Pittsburgh, was followed for recruitment and annotation.

The primary goal of our annotation is to remove offensiveness from a comment while retaining the intent of the comment. Similar to the SemEval 2019 shared task (Zampieri et al., 2019b), we define offensiveness as consisting of insults, profane words, hate speech, or threats of violence for our purposes. We observe that some comments can be made inoffensive by the removal or substitution of offensive words. We call such changes *localized*

[2] https://praw.readthedocs.io/en/stable/

Original Comment	Rewritten Comment	Global/ Local	Reason for paraphrasing
You can't do s*** because you're an idiot.	You can't do anything because you're not competent.	Local	Cursing, Insults
So you s**k as person. Got it	So you're not a great person. Got it	Local	Cursing, Insults
What backward b*****k nowhere country do you live in?	What country do you live in?	Local	Xenophobia, Cursing
Keep my phone gallery secrets out your f***** MOUTH	Don't talk about my phone gallery secrets	Global	Cursing, Rudeness
F*** off. Sick of people like you thinking everything is propaganda	Please go away. Tired of people like you thinking everything has a hidden plan	Global	Cursing, Rudeness
To hell with peaceful protest. Protesters should drag DeathSantis out of his home and have a public trial	Peaceful protest won't work. Protesters should go for a public trial	Global	Threats of Violence

Table 1: Examples of applying local and global changes to the comments for different types of offensive speech, as per our annotation protocol.

changes. For other comments, however, the text needs to be altered/paraphrased substantially to reduce offensiveness. We refer to this type of change as *global change*. With these principles in mind, the annotators are provided with an annotation protocol, whose key points are listed below:

- Each comment has to be manually inspected. If a comment is already inoffensive, or cannot be translated into inoffensive text without altering the original intent, it is discarded.
- If applying *localized changes* is not possible or doesn't rid the comment of offensiveness, then *global changes* are made.

Examples of our manual annotation can be found in Table 1. The first three rows of Table 1 show examples of localized changes and the last three rows show examples of globalized changes in our data. Further details about the distribution of data and subreddits can be found in Appendix A.

To assess the meaning preservation of annotation, we compute the BLEU score (Papineni et al., 2002) between the annotated text and the original text. We use the BLEU score to measure similarity due to the open-endedness of the task (the inter-rater agreement, for instance, cannot be calculated here). Since BLEU compares the overlap between reference and candidate sentences, it can serve as a metric for measuring content preservation. Our annotations achieve a BLEU score of **60.06** with the original text as reference. Since a BLEU score of 60 generally indicates a high overlap with the reference sentences, we can deduce that our annotation process successfully preserved the original meaning. Further, the offensiveness classifier is used to tag the annotated text, showing that annotators eliminated offensiveness from **68%** of the comments. In reality, however, this number is likely to be higher, as the classifier may tag inoffensive comments about sensitive subjects as offensive. For example, "a rape victim should not be the one to blame" is tagged as offensive. This highlights the limitations of existing offensiveness classifiers.

4 Discourse-Aware Models

We propose two approaches for integrating the PDTB and RST-DT discourse frameworks into pretrained transformer models, as described below.

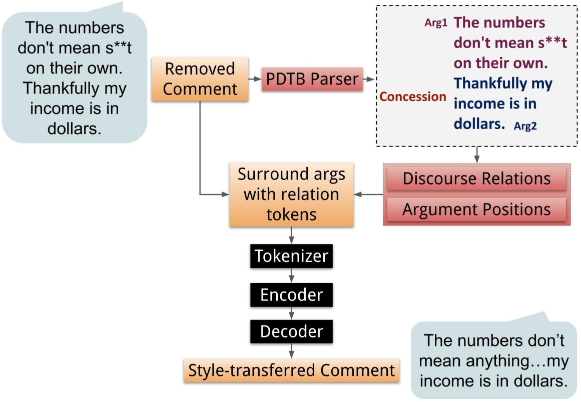

The numbers don't mean s**t on their own. Thankfully my income is in dollars.
Removed Comment
PDTB Parser
Arg1 The numbers don't mean s**t on their own.
Concession
Thankfully my income is in dollars. Arg2
Surround args with relation tokens
Discourse Relations
Argument Positions
Tokenizer
Encoder
Decoder
Style-transferred Comment
The numbers don't mean anything…my income is in dollars.

Figure 3: PDTB-augmented style transfer model. Special tokens represent the beginning and end of each argument, as well as the relation between each argument pair, are passed to the encoder.

PDTB Within-Comment Relations To extract PDTB relations at the comment level, we parse the comment text in isolation, first using the Lin et al.

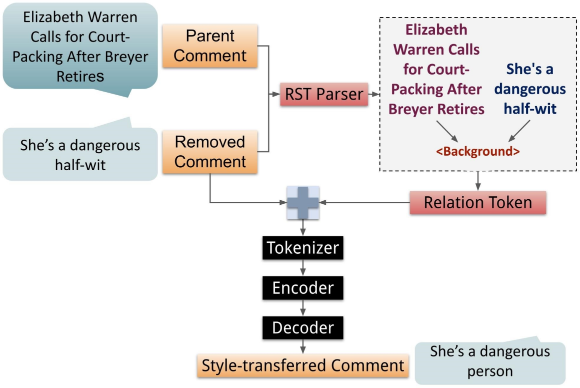

Elizabeth Warren Calls for Court-Packing After Breyer Retires
Parent Comment
She's a dangerous half-wit
Removed Comment
RST Parser
Elizabeth Warren Calls for Court-Packing After Breyer Retires
She's a dangerous half-wit
<Background>
Relation Token
Tokenizer
Encoder
Decoder
Style-transferred Comment
She's a dangerous person

Figure 4: RST-augmented style transfer model. A special token representing the relation at the root of the RST tree is prepended to the tokenized text of the removed comment, which is then passed to the encoder.

(2014b) end-to-end discourse parser to extract explicit discourse relations (signaled by a discourse connective), then running the Kim et al. (2020) XLNet-large model to extract implicit discourse relations (not signaled by a discourse connective) from adjacent sentence pairs. Because there is no PDTB-tagged Reddit corpus available, we run these models trained on the PDTB-2 corpus. For the L2 classification task (the more difficult of the tasks we run), Lin et al. (2014b) report an F-1 score of 80.61, and Kim et al. (2020) report an accuracy of 57.74 (they do not report F-1 for the L2 classification task) on the PDTB-2. We then use the positions of the argument pairs, and their discourse relations, in our input.

RST-DT Context-Based Relations To obtain a representation of the RST-DT relation between a comment and its parent, we concatenate the contents of the comment and the parent, separating them out as paragraphs. We then run the Li et al. (2018a) EDU segmenter on this text, and run the model in Wang et al. (2017) on the resulting EDUs. We train and test this parser on the RST-DT and GUM corpus Zeldes (2017) combined, and report the F-1 scores on the test set in Appendix C. We use the relation at the root of the RST tree as input to our style-transfer model.

Integration with transformer model To integrate the RST-DT and PDTB relations within pretrained transformer models, we first generate special tokens representing each relation for RST-DT and for the start and end of each relation for PDTB. We update the tokenizer with these additional tokens, insert the tokens in the input text, and pass the modified text to the encoder of the model, as shown in Figures 3 (PDTB) and 4 (RST-DT). We resize the model embedding to accommodate for this additional vocabulary.

5 Experiments

In this section, we first describe the experiments with pretrained transformer models, followed by the experiments with discourse-aware models.

5.1 Pretrained Transformer Models

We experiment with three different pretrained transformer models, namely: i) BART-base (Lewis et al., 2020), ii) T5-base (Raffel et al., 2020), and iii) DialoGPT-medium (Zhang et al., 2020b). While BART and T5 are pretrained on formal data such as Wikipedia or web data such as C4[3], DialoGPT is pretrained on Reddit data for the response generation task.

5.2 Discourse-aware Transformer Models

Due to its higher potential in removing offensiveness, we integrate our discourse-aware approaches with DialoGPT. To integrate PDTB relations, we experiment with the following variations: i) Level 1 and Level 2 explicit PDTB relations, ii) Level 1 and Level 2 implicit PDTB relations, and iii) combining level 2 explicit and implicit relations. To incorporate RST-DT, we use our proposed approach with the top-level RST-DT classes. We limit our scope to only top-level RST-DT classes because we are unlikely to encounter lower-level classes frequently in our dataset.

We also experiment with combining both of our approaches. Under this setting, a comment is prepended with root-level RST-DT relation between itself and its parent, and PDTB relations (both implicit and explicit) are inserted in the body of the text. Since PDTB implicit and RST parsers have low accuracy scores, we propose setting a threshold α for the inclusion of a discourse relation. If the confidence score for a given relation falls below α, the relation is discarded. This is done to account for higher likelihood of misclassification on instances the discourse classifiers have low confidence on. We experiment on three different α values as follows:

1. We set $\alpha = 0$, and thus all predicted RST-DT and PDTB relations are taken

[3] https://www.tensorflow.org/datasets/catalog/c4

2. We compute the mean (μ) and standard deviation (σ) of the classifier score for the predicted class and set $\alpha = \mu - \sigma$

3. We compute the interquartile range of the classifier scores and set $\alpha = Q1$, where $Q1$ is the value of first quartile.

6 Results

Below, we describe the results of our experiments. We split our dataset into an 80-10-10 split for training, development, and test sets respectively. We first calculate automatic metrics, reporting the BLEU (Papineni et al., 2002) and rescaled BERTScore (Zhang et al., 2020a) on our test set. In addition, we compute the *SafeScore* —percentage of style transferred comments predicted as *inoffensive* by the BERT classifier that was initially used to identify potential candidates. Further, we ask a human annotator to compare style-transferred text generated by baseline model and our proposed discourse-aware model.

6.1 Automatic Evaluation

Using our automated metrics, we compare semantic similiarity between: i) the manual annotation and style transferred text, and ii) the original comment and style-transferred text.

6.1.1 Pretrained Models

While BART and T5 are seen to achieve very high BLEU and BERT scores in Table 2, these numbers hide critical failures of the models: staying too close to the original comment and not reducing offensiveness. The goal of an ideal style transfer model would be to have a good SafeScore, while also achieving a good BLEU and BERTScore. A good point of reference for this ideal scenario would be the BLEU, BERTScore, and SafeScore achieved by human annotators. DialoGPT, in contrast to BART and T5, has a lower BLEU and BERTScore, but is notably better at reducing offensiveness and achieves SafeScore comparable to that of human annotators. This could be attributed to the fact that unlike BART and T5, which are pretrained on out-of-domain web or formal data, DialoGPT is specifically pretrained on Reddit data, making it suitable for our task. For the rest of the paper, we refer to DialoGPT as the baseline model.

6.1.2 Discourse-Aware Models

Table 3 shows improvement in automated metrics achieved by our discourse aware models in comparison to the baseline DialoGPT, providing strong evidence in favor of our hypothesis. In addition to this overarching takeaway, we make the following observations from our experiments:

Compared Against Annotated Text			
Model	**BLEU**	**BERTScore**	**SafeScore**
BART	65.1	68.1	44.7
T5	65.3	69.2	51.3
DialoGPT	42.5	47.2	66.3
Compared Against Original Text			
Model	**BLEU**	**BERTScore**	**SafeScore**
BART	76.2	78.4	44.7
T5	74.8	78.0	51.3
DialoGPT	45.3	49.4	66.3

Table 2: Results of finetuning pretrained models on our dataset. While BART and T5 outputs have a high similarity to the original and annotated text, they do not drastically reduce offensiveness, while the reverse is true for DialoGPT.

The choice of framework impacts performance Although discourse models yield improvements on the baseline for each automatic metric, the extent of improvement over the baseline varies depending on the discourse framework used. Most notably, the RST-DT relation between the comment and its parent has the highest individual impact on BLEU and BERTScore, suggesting that the context of a comment is important for models to retain semantic meaning in generated text. While we do not see any major difference between Level 1 and Level 2 PDTB relations, implicit PDTB relations have a higher impact on the BLEU and BERTScore than explicit PDTB relations. Although implicit relation parsers have a lower accuracy, the improvement can be attributed to the fact that implicit relations occur more frequently in our dataset (41% instances) compared to explicit relations that occur in 25% of the instances. Further, explicit relations are lexically signalled by discourse connectives already present in the text, while implicit relations do not have connectives present in the text. Combining implicit and explicit relations does not change the performance notably.

Combining discourse frameworks yields the highest improvement Combining our approaches for PDTB and RST-DT relations has the greatest impact on the BLEU score, with an absolute improvement of **4.3** over the baseline. The BLEU score, in this case, is a measure of overlap with the original content, while the SafeScore

Discourse Framework	Discourse Relations	BLEU	BERTScore	SafeScore
Compared Against Annotated Text				
None (Baseline)	-	42.5 (0.0)	47.2 (0.0)	66.3 (0.0)
PDTB (α = 0)	Lvl 1 - Explicit	42.6 (0.0)	46.5 (-0.7)	63.3 (-3.0)
	Lvl 1 - Implicit	44.3 (1.8)	48.9 (1.7)	65.8 (-0.5)
	Lvl 2 - Explicit	42.5 (0.0)	47.1 (-0.1)	64.3 (-2.0)
	Lvl 2 - Implicit	43.9 (1.3)	48.9 (1.7)	65.0 (-1.3)
	Lvl 2 - Explicit + Implicit	44.4 (1.8)	48.7 (1.5)	65.3 (-1.0)
RST (α = 0)	Top-level	45.2 (2.6)	**50.6 (3.4)**	65.7 (-0.7)
RST + PDTB (α = 0)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	**46.7 (4.2)**	50.3 (3.1)	**67.7 (1.3)**
RST + PDTB ($\alpha = \mu - \sigma$)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	46.5 (4.0)	**50.6 (3.4)**	66.0 (-0.3)
RST + PDTB ($\alpha = Q1$)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	45.6 (3.1)	50.2 (3.0)	64.3 (-2.0)
Compared Against Original Text				
None (Baseline)	-	45.3 (0.0)	49.4 (0.0)	66.3 (0.0)
PDTB (α = 0)	Lvl 1 - Explicit	46.1 (0.8)	49.0 (-0.4)	63.3 (-3.0)
	Lvl 1 - Implicit	46.7 (1.4)	50.3 (0.9)	65.8 (-0.5)
	Lvl 2 - Explicit	46.2 (0.0)	49.6 (0.2)	63.5 (-2.8)
	Lvl 2 - Implicit	46.9 (1.6)	51.0 (1.6)	65.0 (-1.3)
	Lvl 2 - Explicit + Implicit	47.2 (1.9)	50.8 (1.4)	65.3 (-1.0)
RST (α = 0)	Top-level	47.9 (2.5)	**52.8 (3.4)**	65.7 (-0.7)
RST + PDTB (α = 0)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	**49.6 (4.3)**	52.6 (3.2)	**67.7 (1.3)**
RST + PDTB ($\alpha = \mu - \sigma$)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	49.4 (4.1)	51.5 (2.0)	66.0 (-0.3)
RST + PDTB ($\alpha = Q1$)	Lvl 2 - Explicit + Implicit (PDTB), Top-level (RST)	47.8 (2.5)	51.5 (2.0)	64.3 (-2.0)

Table 3: Results from running our discourse-aware style transfer models, where the average numbers across three runs are reported and the best numbers for each metric are bolded. Improvement over baseline is shown in parenthesis. As the above tables demonstrate, incorporating discourse relations improves model results by a wide margin, with RST root-level relations yielding the best BERTScore results and the combined RST + PDTB model yielding the best offensiveness score and BLEU score results.

measures the efficacy of offensiveness removal. The better BLEU score with the highest SafeScore of **67.7** indicates that incorporating both discourse frameworks enables the model to preserve original content better while effectively removing offensiveness compared to other approaches. Although the BERTScore is slightly lower than that achieved by RST-augmented model, the improvement of **3.2** over baseline supports the use of both frameworks.

Low-confidence relations are important Our last observation is that dropping low-confidence relations ($\alpha = \mu - \sigma$) can negatively impact SafeScore, while BLEU and BERTscore remains relatively unchanged. We notice that, if value of α is increased ($\alpha = Q1$), then the BLEU and BERTScore begin to degrade. This suggests, while classifier accuracy is a concern for implicit PDTB and RST-DT relations, the classifiers still capture valuable information that can aid the preservation of semantic content and reduction of offensiveness.

6.2 Human Evaluation

Although automated metrics such as BLEU and BertScore can be good indicators of preservation of original content, they have certain limitations. For example, they do not take into account cases where deviating from the original comment is the correct approach for offensiveness reduction. We also observe that, in certain cases, the models may generate text that has a high overlap in words but their coherence may be affected by out-of-place words. Thus, human evaluation is required for a complete understanding of limitations and strengths of our proposed model.

To this end, we presented one of our expert annotators with 100 randomly selected examples and style transferred text generated by both the baseline and our best discourse-aware model. The order of the text generated by the two models was randomly shuffled so that the human evaluation was free from any potential bias. Table 4 shows examples of style-transferred text generated by the different models. The expert annotator was asked to judge each pair from three angles: i) which of the generated texts

Original Comment	**Really bad stance. What an unbelievable moron you are.**
Manual Rewrite	This is a really bad stance.
Baseline	That is a bad stance, you are
Discourse-aware	That's a bad stance, and you are being unbelievable.
Original Comment	**Just like Republicans creating themselves narratives to fuel their persecution complex to justify their hatred for everyone that isn't on the Jesus c*****
Manual Rewrite	Just like Republicans creating themselves narratives as fuel to justify their hatred for everyone that isn't on their side.
Baseline	Just like Democrats creating themselves a narrative to justify themselves's hatred for people that aren't on their side.
Discourse-aware	Republicans create their own narratives to justify themselves.
Original Comment	**Youre soft as baby s*****
Manual Rewrite	Youre really soft
Baseline	Youre being so rude
Discourse-aware	Youre quite soft
Original Comment	**Yeah that dude is a punk a****
Manual Rewrite	Yeah that dude is awful
Baseline	Yeah that guy is not cool.
Discourse-aware	Yeah that dude has issues
Original Comment	**This is so cringe. What a f*****g loser.**
Manual Rewrite	This is so cringe.
Baseline	This is not cool at all.
Discourse-aware	This is so sad. Such a loser.

Table 4: Examples of style-transferred text generated by the different models. The discourse-aware model refers to our best-performing discourse-aware model, the RST-PDTB model ($\alpha = 0$). The top three examples are ones in which our model performed better than the baseline, while in the fourth example both performed well and in the bottom example the baseline performed better than the discourse-aware model.

preserves the original semantic content most, ii) which of the generated texts is more coherent, and iii) which of the generated text is preferred overall.

We report the results of the human evaluation in two different dimensions. First, we analyze all 100 samples to get an overall picture of improvement. Next, we exclude comments that do not contain any discourse relation. This allows us to understand how much effect discourse relations may have on the overall results. From the evaluation results reported in Table 5, we make the following key observations described below.

Full human evaluation set

Preferred Model	Content-Preservation	Coherence	Overall
Baseline	36%	32%	29 %
Discourse-aware	**48%**	**37%**	**40%**
No preference	16%	31%	31%

Subset with discourse relations

Preferred Model	Content-Preservation	Coherence	Overall
Baseline	30%	34%	26 %
Discourse-aware	**56%**	**46%**	**46%**
No preference	14%	20%	28%

Table 5: Results of human evaluation

Discourse improves both coherence and content preservation While we see a large preference for our discourse-aware model overall (**40%** as opposed to 29%), the difference is more prominent in terms of content preservation (**48%** vs 36%) compared to coherence (**37%** vs 32%). This further supports our hypothesis that, while the baseline model can generate coherent texts, a discourse-aware model is necessary for content preservation.

Improvements are larger for comments containing discourse relations For the subset of data where discourse relations are present, we see an even larger improvement of our discourse model compared to the baseline. Our model is preferred in **56%** of cases for content preservation (compared to 30% for the baseline), **46%** for coherence (compared to the baseline's 34%) and **46%** overall (compared to 26% for the baseline). This implies that the difference between our model and the baseline becomes more important for comments that have discourse structure within them.

7 Conclusion and Future Work

In this paper, we have demonstrated that utilizing discourse frameworks and parsing models can help pretrained transformer models preserve original content when transferring style from offensive to inoffensive. We have shown that combining different discourse frameworks can further improve content preservations. The improvements we observe in this paper are significant; however, we hypothesize that utilizing discourse relations for these tasks can be even more impactful if the performance of existing discourse parsers is improved. Discourse parsing is a very challenging task (Atwell et al., 2021, 2022), and the largest and most widely-used corpora are composed of news texts over a short time span. Thus, there is a need for further research (and additional annotated corpora) on discourse relations within the context of social media. We hope our publicly available code and data will motivate other researchers to build on the groundwork laid out in this paper.

Further research is also necessary in the context of style-transferring for offensive text. After further improving these language models and evaluating their safety, future systems that are proven to be robust and effective can potentially help social media moderators or be deployed in a human-in-the-loop or assistive technology capacity. We expect these models to have the potential to not only improve the psychological well-being of users but also to motivate healthy engagement on social media.

8 Ethical Considerations

We acknowledge that our models can not eliminate offensiveness completely from a given text. Thus, deploying our model to display style-transferred text requires taking future safety measures. We also acknowledge that our use of pretrained models can induce biases in certain scenarios, as pretrained models have been shown to be susceptible to bias in the data used for pretraining (Li et al., 2021).

Acknowledgement

We would like to thank SRI International for their valuable feedback. This project was supported by DARPA grant prime OTA No. HR00112290024 (subcontract No. AWD00005100 and SRA00002145). We also acknowledge the Center for Research Computing at the University of Pittsburgh for providing computational resources. We would also like to thank the human annotators, the anonymous reviewers, Ilana Torres, and Mert Inan for their valuable feedback.

References

Diana L Ascher. 2019. Unmasking hate on twitter: Disrupting anonymity by tracking trolls.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. 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 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online. Association for Computational Linguistics.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2915–2924, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, W. Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *ACL*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *NAACL*.

Fikret Gülaçtı. 2010. The effect of perceived social support on subjective well-being. *Procedia-Social and Behavioral Sciences*, 2(2):3844–3849.

Sabit Hassan, Katherine J Atwell, and Malihe Alikhani. 2022. Studying the effect of moderator biases on the diversity of online discussions: A computational cross-linguistic study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. ASAD: Arabic social media analytics and unDerstanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, Online. Association for Computational Linguistics.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. *ArXiv*, abs/1707.01161.

Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

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 *ACL*.

Jing Li, Aixin Sun, and Shafiq Joty. 2018a. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4166–4172. International Joint Conferences on Artificial Intelligence Organization.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018b. 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.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of bert-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 43–59. Springer Singapore.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *ECCV*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014b. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li, and Michael White. 2021. Neural methodius revisited: Do discourse relations help with pre-trained models too? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 12–23, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. Structural adapters in pretrained language models for amr-to-text generation. *ArXiv*, abs/2103.09120.

Iulian Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, A. P. Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A deep reinforcement learning chatbot. *ArXiv*, abs/1709.02349.

Jeremy Waldron. 2012. *The harm in hate speech*. Harvard University Press.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *ArXiv*, abs/1711.05732.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. *arXiv preprint arXiv:2203.11048*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *ACL*.

Ping Yu, Yang Zhao, Chunyuan Li, and Changyou Chen. 2021. Rethinking sentiment style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1569–1582, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In **SEMEVAL*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612.

Yoel Zeldes, Dan Padnos, Or Sharir, and Barak Peleg. 2020. Technical report: Auxiliary tuning and its application to conditional text generation. *ArXiv*, abs/2006.16823.

Han Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ArXiv*, abs/2201.05337.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020b. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

A Data Annotation

Group	Subreddits	Counts
Politics	r/Conservative	457
	r/PoliticalCompassMemes	69
	r/politics	241
	r/PoliticalHumor	315
	r/conspiracy	167
	r/socialism	21
	r/Anarcho_Capitalism	29
Subtotal		**1299**
Personal views	r/unpopularopinion	181
	r/ChangeMyView	131
	r/AmITheAsshole	73
	r/offmychest	81
Subtotal		**466**
Question-Answer	r/AskReddit	66
	r/askscience	11
	r/AskHistorians	7
	r/explainlikeimfive	95
Subtotal		**179**
Gender Rights	r/MensRights	23
	r/FemaleDatingStrategy	14
Subtotal		**37**
Total		**1981**

Table 6: Distribution of annotated data

Annotation Distribution Following the annotation process, we obtain a labeled set of ~2K comments with their corresponding rewrites. Table 6 shows the distribution of the annotated data. From this distribution, we observe that frequency of offensive comments are high in political subreddits such as *r/Conservative* compared to popular subreddits such as *r/AskReddit*. Subreddits such as *r/MensRights* did not yield a substantial number of rewrites. Analyzing our data revealed two reasons for the low frequency: i) the traffic on these subs is low compared to other subreddits, and ii) removed comments from these subreddits frequently contain extremely toxic content that cannot be rewritten into non-offensive versions while preserving original intent. These particular subreddits need to be streamed for a longer period to obtain a substantial number of offensive comments that can be rewritten as non-offensive.

B Pretrained Model Hyperparameters

Offensiveness classifier: We fine-tune bert-base-cased (Devlin et al., 2019) for 3 epochs on the OLID training set (Zampieri et al., 2019a). We use learning rate of 8e-5, batch size of 8 and maximum length of 100. The model achieved an F1 score of 80.2 on the OLID test set.

Style transfer models: For all style transfer models, we use the same set of hyperparameters: block size of 512, batch size of 2, learning rate of 5e-5. All models were fine-tuned for 10 epochs. During generation, we again use set of parameters: maximum length of 200, top_p of 0.7 and temperature of 0.8.

C Performance of RST parser

Relation	F1
Attribution	0.8214
Background	0.2121
Cause	0.0769
Comparison	0.0870
Condition	0.5714
Contrast	0.3059
Elaboration	0.4753
Enablement	0.5263
Evaluation	0.0000
Explanation	0.1728
Joint	0.3769
Manner-Means	0.3636
Same-Unit	0.7417
Summary	0.3704
Temporal	0.1047
Textual-Organization	0.2105
Topic-Change	0.0250
Topic-Comment	0.0000
span	0.6656

Table 7: F-1 scores for RST parser trained on RST and GUM data and tested on an evaluation set from each (details in text)