Towards A Friendly Online Community: An Unsupervised Style Transfer Framework for Profanity Redaction

Minh Tran[†], **Yipeng Zhang**[§], **Mohammad Soleymani**[†]
[†]Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
[§]University of Rochester, Rochester, NY, USA
[†]{mtran,soleymani}@ict.usc.edu
[§]yzh232@u.rochester.edu

Abstract

Offensive and abusive language is a pressing problem on social media platforms. In this work, we propose a method for transforming offensive comments, statements containing profanity or offensive language, into non-offensive ones. We design a RETRIEVE, GENERATE and EDIT unsupervised style transfer pipeline to redact the offensive comments in a word-restricted manner while maintaining a high level of fluency and preserving the content of the original text. We extensively evaluate our method's performance and compare it to previous style transfer models using both automatic metrics and human evaluations. Experimental results show that our method outperforms other models on human evaluations and is the only approach that consistently performs well on all automatic evaluation metrics.

1 Introduction

Despite the undeniably positive impact social media has on facilitating communication, it is also a medium that can be used for abusive behavior. Many social media platforms do not restrict the language users use, leading to an overflow of strong language that might not be appropriate for children (Duggan, 2014; Rieder, 2010). Verbal abuse and cyber-bullying is also a common problem on social media. Such phenomena are harmful to the victims, the online community, and in particular adolescents who are more susceptible and vulnerable in such situations (Patchin and Hinduja, 2010; Pieschl et al., 2015). To mitigate such problems, recent studies have focused on developing machine learning models for detecting hate speech (Davidson et al., 2017; Xiang et al., 2012; Djuric et al., 2015; Waseem and Hovy, 2016; Chen et al., 2012; Xiang et al., 2012; Founta et al., 2019). However, little progress has been made regarding the task of transforming hateful sentences into non-hateful ones, a potential next-step after detecting the hateful content. dos Santos et al. (2018) propose an extension of a basic encoder-decoder architecture by including a collaborative classifier. To the best of our knowledge, this is the only approach dealing with abusive language redaction.

Unsupervised text style transfer is an important area in text generation that has recently received a lot of attention. Generally speaking, text style transfer is the task of rewriting sentences in a source style to a target style while preserving the original sentences as much as possible. In the context of the paper, we define a corpus to be stylistic if every sample in the corpus shares a common style. Most style transfer approaches are developed and validated on bi-stylistic datasets (Shen et al., 2017; Hu et al., 2017; Li et al., 2018; Prabhumoye et al., 2018; Tian et al., 2018; He et al., 2019; Wu et al., 2019), which require stylistic features on both source and target samples. Some common bi-stylistic datasets for text style transfer are the (negative-positive) Yelp restaurant reviews (Shen et al., 2017) & Amazon product reviews (He and McAuley, 2016), (democratic-republican) Political slant (Prabhumoye et al., 2018), (male-female) Gender (Reddy and Knight, 2016) and (factual-romantic-humorous) Caption (Gan et al., 2017). Models training on these datasets are not normally suitable for being trained and validated on uni-stylistic datasets, where only the source or the target set is stylistic (*e.g.*, offensive to normal

Proceedings of the 28th International Conference on Computational Linguistics, pages 2107–2114
Barcelona, Spain (Online), December 8-13, 2020

text). Recently, Madaan et al. (2020) introduce a uni-stylistic Politeness dataset along with a tag-and-generate approach, in which a generator model learns style phrases from the target samples to fill in tagged positions (cannot be generalized to our case where the target sentences are not stylistic).

In this work, we propose a novel Retrieve, Generate and Edit framework to solve the task of transferring offensive sentences into non-offensive ones. For validation, we use three criteria for assessing the performance of our model, namely, content preservation, style transfer accuracy and fluency. We perform an extensive comparison with prior style transfer work on both objective and subjective ratings.

2 Methodology

2.1 Problem Formulation

Given a vocabulary of restricted words V_r and a corpus of labeled sentences $\mathcal{D} = \{(x_1, l_1), \ldots, (x_n, l_n)\}$ where x_i is a sentence and $l_i =$ "offensive" if there exists an offensive word v_i ($v_i \in V_r$) in x_i, otherwise $l_i =$ "non-offensive". For (x_i, l_i) where $l_i =$ "offensive", we re-generate x_i^* such that it does not contain any words from V_r, preserves as much content from x_i as possible, and is grammatical and fluent. Unlike dos Santos et al. (2018), who handle general hateful and offensive content detected by Davidson et al. (2017)'s offensive language and hate speech classifier, we focus our work on profanity removal.

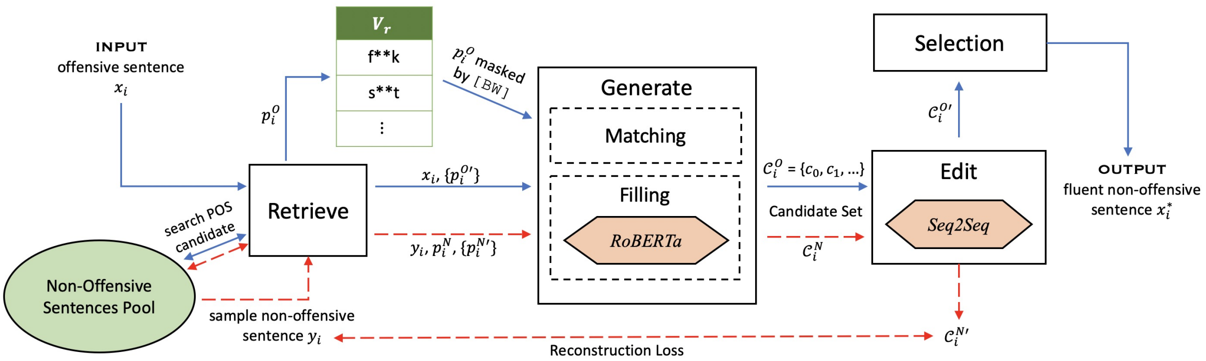

Figure 1: Overview of our Retrieve-Generate-Edit framework. The dotted red arrows denote the steps for training the sequence-to-sequence model while the solid blue ones denote the steps taken during inference. We use superscripts O (offensive) and N (non-offensive) to differentiate the variables.

2.2 Data Collection

We construct the list of 1,580 restricted words V_r from various sources [1,2]. For Corpus $\mathcal{D}$, we extract a total of 12M comments from 2 highly controversial subreddits (6M from each): `r/The_Donald` and `r/politics` from January 2019 to December 2019 using *BigQuery*[3]. We extract sentences that have between 5 and 20 words from the comments. We further remove sentences containing URL, number, email, emoticon, date and time using the *Ekphrasis* text normalization tool (Baziotis et al., 2017). The remaining sentences are then labeled as either "offensive" or "non-offensive", as defined, resulting in 350K "offensive" sentences and 7M "non-offensive" sentences.

2.3 Framework

As shown in Figure 1, our Retrieve, Generate and Edit framework first retrieves possible Part-of-Speech (POS) tagging sequences, which are then used as the templates for generating candidates in the Generate module and corrected by the Edit module.

[1]`https://www.noswearing.com/dictionary`
[2]`https://www.cs.cmu.edu/~biglou/resources/bad-words.txt`
[3]`https://cloud.google.com/bigquery`

Retrieve We first perform POS tagging on both the labelled 350K offensive and 7M non-offensive comments using the *Stanza* POS tagger (Qi et al., 2020). We replace the POS tags of the offensive terms in V_r with a `[BW]` token. Then, given an offensive sentence x_i and its POS sequence p_i, we use the *Lucene* search engine[4](TF-IDF based) to find the set of 10 most similar POS sequences $\{p'_i\}$ that belong to sentences in the non-offensive set.

Generate After getting x_i, p_i and $\{p'_i\}$, the GENERATE module creates a set of sentences $\mathcal{C}_i$ containing no offensive words. The module achieves this by "matching" words in x_i into possible positions in each p'_i to generate new sentences. The positions that are unable to be matched are "filled" by a pretrained language model. The pseudocode for the algorithm can be found in Algorithm 1.

Algorithm 1: Candidate set generation.

```
Input  : x_i, p_i, p'_i, V_r
         T_i - set of unique POS tokens in p_i
         T'_i - set of unique POS tokens in p'_i
         F - pretrained mask-filling model
Output: Set of candidate sentences C_i.
Definition: P^n_k := Value of the k-permutations of n.
T_shared ← T_i ∩ T'_i
c_0 ← [MASK]_1[MASK]_2 ... [MASK]_|p'_i|
C_i ← {c_0}
foreach token t_k in T_shared do
    W_k ← set of words in x_i tagged with t_k
    S_k ← list of t_k's positions in p'_i
    A_k ← list of possible assignments of words in W_k to positions S_k
      ▷ O(max(P^{|W_k|}_{|S_k|}, P^{|S_k|}_{|W_k|}))
    foreach candidate c_j in C_i do
        C_i.remove(c_j)
        foreach assignment a in A_k do
            c'_j ←ASSIGN(a, c_j)
            C_i.add(c'_j)
        end
    end
end
return {F(c_j)}_{j=0,1,...,|C_i|}
```

- **Matching** For each p'_i, we first create a set $\mathcal{T}_{shared}$ of unique shared tokens in p_i and p'_i. We initialize sentence c_0 of length $|p'_i|$ filled with `[MASK]` tokens to store the sentence generated according to p'_i. For a token t_k in $\mathcal{T}_{shared}$, we try to fill all its corresponding positions in c_0 using words in x_i that are tagged with t_k. Suppose there are N words and M positions, then there are at most $max(\frac{N!}{(N-M)!}, \frac{M!}{(M-N)!})$ possible permutations. We find the number to be 9.42 on average for 5K randomly sampled offensive sentences. We add each newly generated sentence c'_j into $\mathcal{C}_i$ and repeat for each t_k on all sentences in $\mathcal{C}_i$ until all their masked positions correspond to tokens not in $\mathcal{T}_{shared}$.
- **Filling** For each resulting candidate sentence in $\mathcal{C}_i$, we use the pretrained RoBERTa-base model (Liu et al., 2019) to fill in remaining `[MASK]` tokens. To enhance content preservation, we insert the original sentence x_i before each of the generated sentences with a `[SEP]` token in between. We replace each `[SEP]` token with the most probable word predicted by RoBERTa that is not in V_r. The unmasked sentences after `[SEP]` are the outputs of the GENERATE module.

Edit We use an EDIT module to correct the problems of the output sentences from the GENERATE module, mostly related to wrong word orderings due to the permutation generation in the MATCHING

[4]https://lucene.apache.org/core/

step or low fluency due to a bad retrieved POS sequence from the RETRIEVE module. We first randomly sample 60K English-only non-offensive sentences and apply the RETRIEVE and GENERATE modules on the chosen sentences (dotted red arrows in Figure 1). In the RETRIEVE module, we retrieve POS sequences $\{p_i^{N\prime}\}$ from the non-offensive set and drop the first retrieved sequence, which is the original query sequence y_i itself. We then form a parallel corpus using the generated candidates $\mathcal{C}_i^N$ as the source dataset while having the original non-offensive sentences as the target dataset, resulting in 780K source-target pairs. In this study, we finetune the pretrained T5-small model (Raffel et al., 2019) as our editing sequence-to-sequence model using the generated parallel corpus. We call the edited candidate set $\mathcal{C}_i'$.

Selection We add a SELECTION module to select the candidate of highest quality x_i^* from $\mathcal{C}_i'$. We first remove any candidate with words in V_r. Then, each generated candidate is assigned a content preservation score (BLEU score (Papineni et al., 2002) between the source and the candidate sentences) and a fluency score (perplexity estimated by the pretrained GPT-2 model with 117M parameters[5] (Radford et al., 2019)). The content preservation and fluency scores are then normalized to $[0, 1]$ by *MinMaxScaler*. The candidate with the highest sum of content preservation and fluency scores is chosen.

3 Experimental Results

3.1 Baselines

We compare our framework (`R+G+S` and `R+G+E+S`)[6] against 8 existing style transfer methods. These methods are: cross-alignment `CA` (Shen et al., 2017), back-translation `BT` (Prabhumoye et al., 2018), delete-only `DL` and delete-retrieve-generate `DRG` (Li et al., 2018), mask-and-infill `MLM` (Wu et al., 2019), auto-encoder with POS information preservation constraint `AEC` (Tian et al., 2018), deep latent sequence model `DLS` (He et al., 2019) and the tag-and-generate model `TG` (Madaan et al., 2020). We also compare our method with the removal approach `REM`, which simply removes offensive terms from sentences.

For all baselines methods, we replicate the experimental setups described in their papers. Since some of the baseline models' performance are susceptible to unbalanced classes during training (Li et al., 2018; Wu et al., 2019; Tian et al., 2018), we subsample the non-offensive sentences from 7M to 350K sentences, resulting in a balanced dataset. We then split the offensive and non-offensive datasets into train (320K), validation (25K) and test (5K) sets. Implementation details can be found in Appendix A.

3.2 Evaluations

Automatic Evaluations Following most prior studies on text style transfer, we use 3 criteria to evaluate the generated outputs of the models: content preservation, style transfer accuracy and fluency. For content preservation, we report the BLEU-self (BL) (Papineni et al., 2002), ROUGE (RG) (Lin, 2004) and METEOR (MT) (Denkowski and Lavie, 2011). We calculate the style transfer accuracy (Acc.) as the percentage of generated sentences not containing any words in V_r. For fluency, we use the average perplexity (PPL) of generated sentences calculated by the pretrained GPT-2 model (Radford et al., 2019).

Model	BL ↑	RG ↑	MT ↑	Acc. ↑	PPL ↓
CA	18.3	36.2	11.9	65.0	747.7
MLM	49.7	63.3	40.8	65.5	798.6
AEC	46.7	56.3	25.9	90.2	3470.6
BT	8.5	21.3	9.3	95.2	488.5
DLS	30.9	48.8	17.9	99.1	445.9
R+G+S	51.8	67.7	41.5	100.0	674.9
R+G+E+S	47.4	57.7	33.9	99.6	448.7
REM	81.3	87.9	49.0	100.0	1259.8

Table 1: Automatic evaluation results. For each metric, we mark the 3 best/worst-performing models in green/red. The average perplexity of the original sentences is 458.1.

We show the performances of methods that have at least 60% accuracy in Table 1, while reporting the remaining ones in Appendix B. Our models are the only ones that consistently perform among the

[5]`https://huggingface.co/gpt2`

[6]RETRIEVE, GENERATE, [EDIT] and SELECTION. The EDIT module can be skipped.

top in all 3 criteria. The perplexity of `R+G+E+S` is lower than the perplexity of `R+G+S` by 226 points, suggesting the effectiveness of the trained sequence-to-sequence model to edit the output candidates from the GENERATE module.

Although we do not compare the performance of our framework with (dos Santos et al., 2018), we use the same set of evaluation metrics reported in their work. On a training dataset of size 224K offensive sentences and 7M non-offensive Reddit sentences, dos Santos et al. (2018) report a content preservation score, proposed by Fu et al. (2018), of 0.933, a style transfer accuracy of 99.54% and a worse perplexity than `CA`'s outputs. For reference, our best performing model, `R+G+E+S`, achieves a Fu et al. (2018)'s content preservation score of 0.965, a style transfer accuracy of 99.6% and a better perplexity than `CA`.

Human Evaluations We ask 3 unbiased human judges to rate the outputs of our models, as well as `MLM` and `DLS`, which are the 4 best performing models according to the automatic evaluation metrics. Following Li et al. (2018), the annotators judge the generated sentences on content preservation (CP) and grammaticality (Gra.) on a scale from 1 to 5. From 5K offensive sentences in the test set, we randomly sample 100 offensive sentences and ask the annotators to rate the generated outputs of the 4 models on these chosen sentences. We report the style transfer success rate (Succ.) for each method, which is calculated as the number of sentences that do not contain any words from V_r and receive an average CP and Gra. scores of at least 4. Table 2 shows the results of the manual evaluations, which demonstrates a significantly higher Succ. score of `R+G+S` and `R+G+E+S` in comparison with previously published models. Some generated samples of the 4 methods are available in Appendix C.

Model	**CP ↑**	**Gra. ↑**	**Acc. ↑**	**Succ. ↑**
DLS	1.947	4.037	99%	7%
MLM	3.157	**4.383**	73%	18%
R+G+S	**3.650**	3.840	**100**%	40%
R+G+E+S	3.567	4.077	**100**%	**46**%

Table 2: Human evaluation results.

4 Conclusion

In this paper, we propose a novel RETRIEVE, GENERATE and EDIT text style transfer framework that redacts offensive comments on social media in a word-restricted manner. The experimental results on both automatic metrics and manual evaluations demonstrate the strong performance of our method over prior models for the given task. For future work, we envision the possibility of extending the framework by automatically detecting the restricted vocabulary set V_r. Such ability would enable the framework to be a robust style transfer method that is applicable to both uni-stylistic and bi-stylistic datasets.

Acknowledgements

Research was in-part sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 747–754.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30.

Cicero 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.

Maeve Duggan. 2014. *Online harassment*. Pew Research Center.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Rroceedings of the 25th International Conference on World Wide Web*, pages 507–517.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

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-Volume 70*, pages 1587–1596.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 arXiv:1907.11692*.

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, July. 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 *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Justin W Patchin and Sameer Hinduja. 2010. Cyberbullying and self-esteem. *Journal of school health*, 80(12):614–621.

Stephanie Pieschl, Christina Kuhlmann, and Torsten Porsch. 2015. Beware of publicity! perceived distress of negative cyber incidents and implications for defining cyberbullying. *Journal of School Violence*, 14(1):111–132.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Rem Rieder. 2010. No comment: it's time for news sites to stop allowing anonymous online comments. *American journalism review*, 32(2):2–3.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: applying masked language model to sentiment transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5271–5277. AAAI Press.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Appendix A. Implementation Details

We finetune the T5-small model in the EDIT module with a learning rate of $1e^{-4}$, the Adam optimizer, cross-entropy loss function and a training batch size of 256 for 3 epochs. We set the max length of input/output sequences to be 30 and the beam size to be 5.

Appendix B. Full Automatic Evaluation Results

Model	**BL** ↑	**RG** ↑	**MT** ↑	**FuCP** ↑	**Acc.** ↑	**PPL** ↓
TG (Madaan et al., 2020)	66.1	76.3	45.4	0.960	23.4	3378.5
DL (Li et al., 2018)	51.8	63.4	30.1	0.931	56.8	811.0
DRG (Li et al., 2018)	47.9	59.6	28.3	0.927	57.2	1113.4
CA (Shen et al., 2017)	18.3	36.2	11.9	0.907	65.0	747.7
MLM (Wu et al., 2019)	49.7	63.3	40.8	0.983	65.5	798.6
AEC (Tian et al., 2018)	46.7	56.3	25.9	0.912	90.2	3470.6
BT (Prabhumoye et al., 2018)	8.5	21.3	9.3	0.900	95.2	488.5
DLS (He et al., 2019)	30.9	48.8	17.9	0.915	99.1	445.9
R+G+S (Ours)	51.8	67.7	41.5	0.977	100.0	674.9
R+G+E+S (Ours)	47.4	57.7	33.9	0.965	99.6	448.7
REM (remove only)	81.3	87.9	49.0	0.986	100.0	1259.8

Table 1: Complete Automatic evaluation results. **FuCP** refers to the content preservation metric proposed by Fu et al. (2018).

Appendix C. Example Outputs

Model	**Generated Output**
Original	reap what you sow a*s clowns.
DLS	except what you believe us.
MLM	this is what you sow and amazing amazing.
R+G+S	reap what you sow clowns.
R+G+E+S	reap what you sow.
Original	the benghazi b*tch is going down.
DLS	the president cruz is going down.
MLM	the benghazi wall is going down.
R+G+S	the benghazi is down going down.
R+G+E+S	the benghazi girl is going down.
Original	put your head back up your a*s and stfu.
DLS	put your head back up your own and then.
MLM	stuff your smart back up your a*s and amazing great.
R+G+S	put your head back up your back, and stfu.
R+G+E+S	stfu and put your head back up.
Original	the w*ore of babylon speaks.
DLS	the state of least run.
MLM	the house of babylon speaks.
R+G+S	babylon speaks the woman of babylon speaks.
R+G+E+S	the babylon speaks of it.
Original	you obviously talking through your b*tt because you lack all sense of having a brain.
DLS	you obviously talking on your own words because you want a lot of free.
MLM	you keep talking through your teeth because you lost all sense of having a brain.
R+G+S	you're obviously talking all sense having through you lack all sense of a brain.
R+G+E+S	you obviously lack all sense of having a brain through your mouth when talking.
Original	no one gives a d*mn about what your platform is because it has no merit.
DLS	no one gives a about what your country is because it is no longer.
MLM	no one gives a flip about what your platform is because it has no merit.
R+G+S	one cares about what one is doing because it has no merit.
R+G+E+S	no one cares about what he is doing because it has no merit for.
Original	i have no sympathy for that b*tch and i never will.
DLS	i have no idea for that it would never will be.
MLM	i have tremendous sympathy for that b*tch and i always will.
R+G+S	i have no sympathy for i will and never.
R+G+E+S	i have no sympathy for you and will never.
Original	war is h*ll and he deserves it.
DLS	war is cruz and he did it.
MLM	war is real and he knows it.
R+G+S	It is and he deserves it.
R+G+E+S	it is war and he deserves it.

Table 2: Example outputs from our framework, DLS and MLM