Style Transfer: Saving the world from Abusive Speech

Arpan Mangal
IIT Delhi
cs1160321@iitd.ac.in

Deepanshu Jindal
IIT Delhi
cs1160312@iitd.ac.in

Abstract

Hate speech and offensive language on social platforms is rampant these days, creating an unhealthy atmosphere and *"dehumanising effect"*[1]. A lot of work has been done in hate speech detection but limited work has been done to tackle it by transforming offensive content to non-offensive content. We will try to employ a Style Transfer based technique to tackle this menace by rephrasing sentences/tweets to make them non-offensive.

1 Introduction

1.1 Hate speech

Hate speech is a growing problem, with abusive and controversial texts and comments often trending on social media platforms causing hurt feelings for the target groups and an unhealthy atmosphere overall. Recent reports have demonstrated that abusive speech severe real world consequences and digital hatred is now spilling over into the real world. A strong correlation has been shown between anti-minority sentiment on social media and the incidents of violent crimes against such groups(Ant). The high impact and prevalence of hate and abusive speech have been identified as important problem which need to be addressed.

1.2 Previous work

Significant work has been done to detect and filter out the hate speech (Davidson, 2017; Founta, 2018). However, it is not sufficient to just filter them out, maybe we can do better by transforming the offensive content into non-offensive content while still being comprehensible in a somewhat toned down and polite version. This would allow us to not only notify a user who is trying to post offensive content but also allow us to prompt the user with some suggestions to town down the message content to maintain the community standards of the platform.

To counter this problem of hate speech we use style transfer techniques, similar to those used by (Shrimai et al., 2018) to modify the stylistic aspects of text responsible for the offensive nature. Previously, similar work has been done by (Cicero, 2018)[2]. We provide a simple set of techniques to transform abusive speech into a more polite form. Moreover the techniques that we develop are general enough to generalize to any form of style transfer, and provide an efficient and easy to implement baseline method to achieve style transfer.

1.3 Style Tranfer

In context of NLP, style transfer is the technique of adding some specific stylistic properties to the text while preserving the overall structure and meaning of the sentence. For example

> *The food was great but the staff was rude.* $\rightarrow$ *The food was great and the staff was friendly.*

shows *sentiment* transfer while preserving other attributes of sentence.

Excellent work in Style Transfer has been done by (Shrimai et al., 2018) (Fig. 1). They first learn a latent representation of the input sentence by using a back translation model in order to preserve the meaning of the sentence while reducing stylistic properties. Thereafter, they use trained decoder over these latent representations to output the text with certain stylistic properties. The decoder is

[1] https://impakter.com/hurt-feelings-real-danger-hate-speech/

[2] However the code for this work was not publicly available

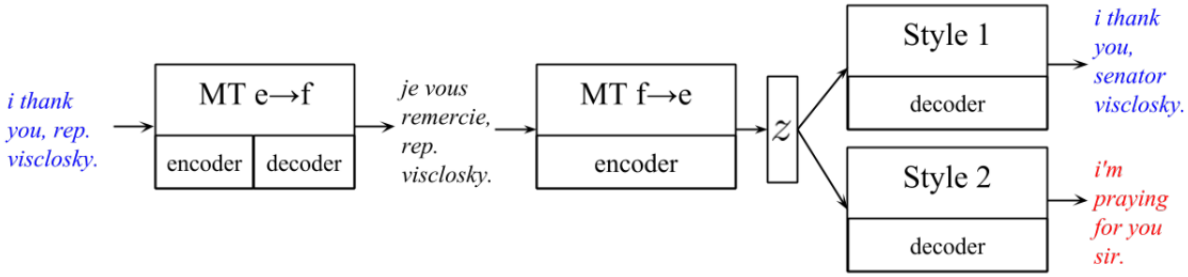

Figure 1: (Shrimai et al., 2018): *Style transfer pipeline: to rephrase a sentence and reduce its stylistic characteristics, the sentence is back-translated. Then, separate style-specific generators are used for style transfer*

trained using feedback from a pre-trained classifier trained to classify the text based on those given stylistic properties. They have trained their model to work with gender, political slant and sentiment based stylistic properties.

2 Datasets and Resources

We work with dataset provided by (Cicero, 2018) which they used for training and evaluation of their model. The dataset has been extracted from reddit and twitter using techniques described by (Henderson et al., 2017) and (Serban et al., 2017).

Apart from the dataset, we use a lexicon of hate words made available by (Davidson et al., 2017).

3 Evaluation Metrics

There are two evaluation metric which we will be focusing on.

1. **Classification Accuracy**: Percentage of times when we are able to convince the classifier that a hate speech text passed through our model is indeed non hate.

2. **Content Preservation**: Gives us measure of how similar the sentence is to original sentence in meaning. For this we use human evaluation.

4 Experiments

4.1 Using Shrimai's model

We train models provided by (Shrimai et al., 2018) over our hate speech dataset.

1. Firstly we preprocess the raw hate speech data to convert sentences like *"i don 't get it..."* to *"i do not get it..."*.

Original (Hate)	Non Hate
that shit isn't on me.	that's what the hell is about me.
i don't get it, who on earth believes shit like that?	i don't know, who would have been on the same thing like this?
less useless meetings and more actually getting shit done, i suppose	reversals? crustaceans and i'm not going to be, i'm sure.

Table 1: Results on hate-speech dataset using Shrimai's model

2. Then we train a CNN classifier to predict hate vs. no-hate style.

3. The English dataset is then translated into French using Shrimai's English-French MT system.

4. Next, the back-translation model is trained, and finally evaluated on test sentences

Results for this are shown in Table 1. The converted sentences gave very high accuracy, comparable to (Cicero, 2018). However the results didn't seem very encouraging in terms of fluency and meaning preservation.

4.2 Lexicon based approach

We also observed that simply removing certain hate words from hate sentences makes them get classified as non-hate more than 90% of the times. This motivates us to try a more comprehensive lexicon based approach to filter out common hate words. We use the lexicon developed by (Davidson et al., 2017) for our task. The lexicon contains

close to 1600 seed words which are used to generate close to 8000 words with polarity scores. We pick the top 2000 words and their common misspellings as the lexicon which we filter out from the hate sentences.

To fill in the gaps in the sentences left out by the removal of these hate words we try following approaches:

4.2.1 CBOW Model

We use Google word2vec embeddings with a CBOW model to predict words to fill in the blanks. However we observe that very often the CBOW model predicts the exact or similar hate words which we removed, which shows the percolation of hate speech into word embeddings as well. (Table 4)

Also many a times simply removing an abusive word is the right way to go.

4.2.2 Backtranslation[3]

(Shrimai et al., 2018) used the approach of back-translation to lose out the original stylistic attributes, preserve latent meaning and finally generating the back-translated sentence in desired style.

We employed a similar approach, by translating the non-fluent sentence (generated after removal of hate words) into a different language, hypothesizing that doing so will help preserve the latent meaning of sentence and back translation to English will smoothen out the discontinuities.

We also observe that back-translation is able to handle subtle hate to some extent as subtle hate gets lost in back-translation.

We also explore Shrimai's hypothesis that "back-translation using a more distant pivot language will help in getting rid of stylistic property to greater extent with the tradeoff of lower quality and fluency of back-translated sentences". We tried our experiments with three completely far off languages – French, Russian and Hindi.

(Li et al., 2018) in their paper on Sentiment Style Transfer use similar technique of deleting words and phrases expressing the original style of the sentence, retrieving new words and phrases of the desired style and use a neural model to fluently combine these.

4.2.3 Results

Using the back-translation technique we got results at par with (Cicero, 2018) (Table 2), in terms of classifier accuracy, fluency and content preservation (Table 5)

[3]Our full code is available here.

Dataset	Shen 2017 (Previous SOTA)	Cicero 2018 (Current SOTA)	Ours
Reddit Hate Speech	87.66 %	**99.54** %	98.60 %

Table 2: Classifier Prediction results for No Hate

5 Conclusion

We provide an interpretable and easy-to-implement style transfer technique to counter the problem of hate speech on online social platforms. Such techniques can be used to automatically convert the hateful content (tweets, comments, and so on) on social chatting platforms like Twitter and Facebook, or communities like StackOverflow.

Also through our experiments with CBOW model for gap filling we have effectively shown how hate speech has percolated into word2vec embeddings trained on news corpus. Such insights can help in understanding some biases in NLP tasks.

6 Proposing a new baseline for style transfer

The techniques that we propose are fairly general and can be a quick baseline to set in any style transfer setting. For lexicon generation we require around a couple hundred of labelled seed words which can be used to induce a large lexicon of relevant words for the style transfer to work upon by using corpus of unlabelled data. Thereafter, we can follow a similar technique of removing words and then filling up with word embeddings of the form $embed(word) + embed(style1) - embed(style2)$.

For example suppose we want to do style transfer for male to female. We induce a lexicon of male and female centric words. Then we filter out male words and fill in the blanks with $embed(word) + embed(man) - embed(woman)$.

Clearly this is not optimal solution but is a quick hack that will work in many practical situations with access to limited labelled data.

References

In germany, online hate speech has real-world consequences.

Igor Melnyk Inkit Padhi Cicero, Nogueira dos Santos. 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)*. Association for Computational Linguistics.

Michael Macy Ingmar Weber Davidson, Thomas Dana Warmsley. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

N. Kourtellis J. Blackburn A. Vakali I. Leontiadis. Founta, D. Chatzakou. 2018. A unified deep learning architecture for abuse detection.

Peter Henderson, Koustuv Sinha, Nicolas Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2017. Ethical challenges in data-driven dialogue systems.

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.

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

Prabhumoye Shrimai, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proc. ACL*.

Original (Female style)	Pretrained Model	Results from training
wonderful wine selection and good service.	great selection of food and good service	pretty good place and good service.
we couldnt even flag him down	we can even get the drapeau	we havent go to this place
i will never order flowers through cactus flowers again	i will never be coming back to try some of the hacked in town .	wow i wouldn't be going back for this place !

Table 3: Results on Shrimai's gender dataset using Shrimai's model (female → male)

Original Sentence:	are you fucking kidding me		
Top 3 Candidates:	fucking 0.92	fuckin 0.81	fuck 0.8
Transformed Sentence:	are you fucking kidding me		
Original Sentence:	i hope this happens so old people quit voting for these jack asses		
Top 3 Candidates:	voting 0.70	for 0.67	vote 0.60
Transformed Sentence:	i hope this happens so old people quit voting for these voting		
Original Sentence:	except when that bitch bites the hand that feeds it		
Top 3 Candidates:	bitch 0.90	bitches 0.71	bastard 0.62
Transformed Sentence:	except when that bitch bites the hand that feeds it		

Table 4: Using CBOW to predict suitable words in blanks

Original	Shen et al.	Cicero et al.	Ours
for fuck sake , first world problems are the worst	for the money , are one different countries	for hell sake , first world problems are the worst	for the sake of the world's first problems are the worst
what a fucking circus this is .	what a this sub is bipartisan .	what a big circus this is .	what a circus it is.
i hope they pay out the ass , fraudulent or no .	i hope the work , we out the UNK and no .	i hope they pay out the state , fraudulent or no .	I hope they pay for it, fraudulent or not.
what big century are you living in ?	life is so big cheap to some people .	you re big pathetic .	in which great century do you live

Table 5: Final results on Cicero's Hate-Dataset