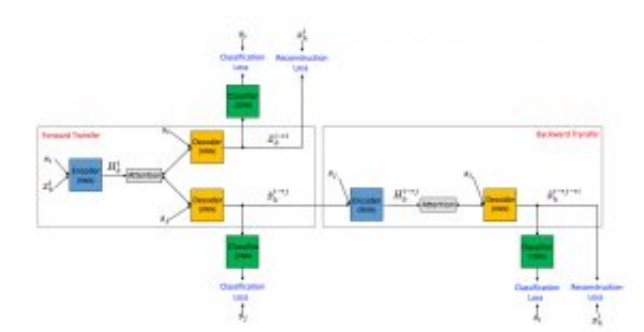
Fighting offensive language on social media with unsupervised text style transfer

July 26 2018, by Cicero Nogueira Dos Santos, Igor Melnyk, Inkit Padhi



Proposed framework of a neural text style transfer algorithm using non-parallel data. Credit: IBM

Online social media has become one of the most important ways to communicate and exchange ideas. Unfortunately, the discourse is often crippled by abusive language that can have damaging effects on social media users. For instance, a recent survey by YouGov.uk discovered that, among the information employers can find online about job candidates, aggressive or offensive language is the most professionally damaging social media activity. Online social media networks normally deal with the offensive language problem by simply filtering out a post when it is flagged as offensive.

In the paper "Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer," which was presented in the 56th

Annual Meeting of the Association for Computational Linguistics (ACL 2018), we introduce a completely new approach to tackle this problem. Our approach uses unsupervised text style transfer to translate offensive sentences into corresponding non-offensive forms. To the best of our knowledge, all previous work addressing the problem of offensive language on social media has focused on text classification only. Those methods can thus be used mainly to flag and filter out the offensive content, but our proposed approach goes one step forward and produces an alternative non-offensive version of the content. This has two potential benefits for users of social media. For those users who plan to post an offensive message, receiving an alert that the content is offensive and will be blocked, along with a more polite version of the message that can be posted, might encourage them to change their minds and avoid the profanity. Additionally, for users consuming online content, this allows them to still see and understand the message but in a non-offensive and polite tone.

An architecture for replacing offensive language

Our method is based on the now popular encoder-decoder neural network architecture, which is the state-of-the-art approach for machine translation. In machine translation, the training of encoder-decoder neural network assumes the existence of a "Rosetta Stone" where the same text is written in both the source and target languages. This paired data enables developers to easily determine whether a system translates correctly and therefore train an encoder-decoder system to do well. Unfortunately, unlike machine translation, as far as we know, there exists no dataset of paired data available for the case of offensive to non-offensive sentences. Moreover, the transferred text must use a vocabulary that is common in a particular application domain. Therefore, unsupervised methods that do not use paired data are needed to perform this task.

	Reddit	Twitter
Original	*for f**k sake , first world problems are the worst*	i 'm back bitc**s ! ! !
(Shen et al., 2017)	for the money , are one different countries	i 'm back ! ! !
Ours	for hell sake , first world problems are the worst	i 'm back bruh ! ! !
Original	*what a f**king circus this is .*	*lol damn imy fake as* lol*
(Shen et al., 2017)	what a this sub is bipartisan .	lol damn imy sis lol
Ours	what a big circus this is .	lol dude imy fake face lol
Original	*i hope they pay out the as* , fraudulent or no .*	*bros before hoes*
(Shen et al., 2017)	i hope the work , we out the UNK and no .	club tomorrow
Ours	i hope they pay out the state , fraudulent or no .	bros before money

We proposed an unsupervised text style transfer approach made up of three main components, each given a separate task during training. One (an RNN encoder) parses an offensive sentence and compresses the most relevant information into a real valued vector. This is read by another component (an RNN decoder), which generates a new sentence that is the translated version of the original one. The translated sentence is then evaluated by the third component (a CNN classifier) to identify whether the output has been correctly translated from the offensive style into non-offensive one. Additionally, the generated sentence is also "back-translated" from non-offensive to offensive and compared to the original sentence to check if the content was preserved. If the results of any of the above evaluations contain errors, the system is adjusted accordingly. The encoder and decoder are also, in parallel, trained using an autoencoding setup where the objective consists in reconstructing the input sentence. We also use the attention mechanism which helps to ensure content preservation. Our main contribution in terms of architecture is the combined use of a collaborative classifier, attention, and back-transfer.

Translating offensive language

We tested our proposed method using data from two popular social media networks: Twitter and Reddit. We created datasets of offensive

and non-offensive texts by classifying approximately 10 million posts using an offensive language classifier proposed by Davidson et al. (2017). The following table shows examples of original offensive sentences and the non-offensive translations generated by a text style transfer method proposed by Shen et al. (2017) and by our approach. Our system demonstrated better performance at translating offensive sentences into non-offensive ones while preserving the overall content but sometimes produces odd sentences.

This work is a first step in the direction of a new promising approach for fighting abusive posts on social media. Unsupervised text style transfer is a research area that has just started to see some promising results. Our work is a good proof of concept that current unsupervised text style transfer methods can be applied to useful tasks. However, it is important to note that current unsupervised text style transfer approaches can only handle well the cases where the offensive language problem is lexical (such as the examples shown in the table) and can be solved by changing or removing a few words. The models we used will not be effective in cases of implicit bias where ordinarily inoffensive words are used offensively.

We believe that improved versions of the proposed method, together with the use of much larger volumes of training data, will be able to cope with other abusive posts such as posts containing hate speech, racism, and sexism. We imagine that our method could be used to improve conversational AI, by ensuring that chatbots that learn by interacting with users online will not later reproduce offensive language and hate speech. Parental control is another potential use of the proposed system.

More information: Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. arxiv.org/abs/1805.07685