Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer

Cicero Nogueira dos Santos* IBM Research T.J. Watson Research Center cicerons@us.ibm.bom

Igor Melnyk* IBM Research T.J. Watson Research Center igor.melnyk@ibm.com

Inkit Padhi* IBM Watson T.J. Watson Research Center inkit.padhi@ibm.com

Abstract

We introduce a new approach to tackle the problem of offensive language in online social media. Our approach uses unsupervised text style transfer to *translate* offensive sentences into non-offensive ones. We propose a new method for training encoder-decoders using non-parallel data that combines a collaborative classifier, attention and the cycle consistency loss. Experimental results on data from Twitter and Reddit show that our method outperforms a state-of-the-art text style transfer system in two out of three quantitative metrics and produces reliable non-offensive transferred sentences.

1 Introduction

The use of offensive language is a common problem of abusive behavior on online social media networks. Various work in the past have attacked this problem by using different machine learning models to detect abusive behavior (Xiang et al., 2012; Warner and Hirschberg, 2012; Kwok and Wang, 2013; Wang et al., 2014; Nobata et al., 2016; Burnap and Williams, 2015; Davidson et al., 2017; Founta et al., 2018). Most of these work follow the assumption that it is enough to filter out the entire offensive post. However, a user that is consuming some online content may not want an entirely filtered out message but instead have it in a style that is non-offensive and still be able to comprehend it in a polite tone. On the other hand, for those users who plan to post an offensive message, if one could not only alert that a content is offensive and will be blocked, but also offer a polite version of the message that can be posted, this could encourage many users to change their mind and avoid the profanity.

In this work we introduce a new way to deal with the problem of offensive language on social media. Our approach consists on using style transfer techniques to *translate* offensive sentences into non-offensive ones. A simple encoder-decoder with attention (Bahdanau et al., 2014) would be enough to create a reasonable translator if a large parallel corpus is available. However, unlike machine translation, to the best of our knowledge, there exists no dataset of parallel data available for the case of offensive to non-offensive language. Moreover, it is important that the transferred text uses a vocabulary that is common in a particular application domain. Therefore, unsupervised methods that do not use parallel data are needed to perform this task.

We propose a method to perform text style transfer addressing two main challenges arising when using non-parallel data in the encoder-decoder framework: (a) there is no straightforward way to train the encoder-decoder because we cannot use maximum likelihood estimation on the transferred text due to lack of ground truth; (b) it is difficult to preserve content while transferring the input to a new style. We address (a) using a single collaborative classifier, as an alternative to commonly used adversarial discriminators, e.g., as in (Shen et al., 2017). We approach (b) by using the attention mechanism combined with a cycle consistency loss.

In this work we also introduce two benchmark datasets for the task of transferring offensive to non-offensive text that are based on data from two popular social media networks: Twitter and Reddit. We compare our method to the approach of Shen et al. (2017) using three quantitative metrics: classification accuracy, content preservation and perplexity. Additionally, some qualitative results are also presented with a brief error analysis.

*Equal contribution.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 189–194
Melbourne, Australia, July 15 - 20, 2018.

2 Method

We assume access to a text dataset consisting of two non-parallel corpora $X = X_0 \cup X_1$ with different style values s_0 and s_1 (offensive and non-offensive) of a total of $N = m+n$ sentences, where $|X_0| = m$ and $|X_1| = n$. We denote a randomly sampled sentence k of style s_i from X as x_k^i, for $k \in 1, \ldots, N$ and $i \in \{0, 1\}$. A natural approach to perform text style transfer is to use a regular encoder-decoder network. However, the training of such network would require parallel data. Since in this work we consider a problem of unsupervised style transfer on non-parallel data, we propose to extend the basic encoder-decoder by introducing a collaborative classifier and a set of specialized loss functions that enable the training on such data. Figure 1 shows an overview of the proposed style transfer approach. Note that for clarity, in Figure 1 we have used multiple boxes to show encoder, decoder and classifier, the actual model contains a single encoder and decoder, and one classifier.

As can be seen from Figure 1, the encoder (a GRU RNN, $E(x_k^i, s_i) = H_k^i$) takes as input a sentence x_k^i together with its style label s_i, and outputs H_k^i, a sequence of hidden states. The decoder/generator (also a GRU RNN, $G(H_k^i, s_j) = \hat{x}_k^{i \to j}$ for $i, j \in 0, 1$) takes as input the previously computed H_k^i and a desired style label s_j and outputs a sentence $\hat{x}_k^{i \to j}$, which is the original sentence but transferred from style s_i to style s_j. The hidden states H_k^i are used by the decoder in the attention mechanism (Bahdanau et al., 2014), and in general can improve the quality of the decoded sentence. When $i = j$, the decoded sentence $\hat{x}_k^{i \to i}$ is in its original style s_i (top part of Figure 1); for $i \neq j$, the decoded/transferred sentence $\hat{x}_k^{i \to j}$ is in a different style s_j (bottom part of Figure1). Denote all transferred sentences as $\hat{X} = \{\hat{x}_k^{i \to j} \mid i \neq j, k = 1, \ldots, N\}$. The classifier (a CNN), then takes as input the decoded sentences and outputs a probability distribution over the style labels, i.e., $C(\hat{x}_k^{i \to j}) = p_C(s_j | \hat{x}_k^{i \to j})$ (see Eq. (2)). By using the collaborative classifier our goal is to produce a training signal that indicates the effectiveness of the current decoder on transferring a sentence to a given style.

Note that the top branch of Figure 1 can be considered as an auto-encoder and therefore we can enforce the closeness between $\hat{x}_k^{i \to i}$ and x_k^i by using a standard cross-entropy loss (see Eq. (1)). However, for the bottom branch, once we transferred X to $\hat{X}$ (forward-transfer step), due to the lack of parallel data, we cannot use the same approach. For this purpose, we propose to transfer $\hat{X}$ back to X (back-transfer step) and compute the reconstruction loss between $\hat{x}_k^{i \to j \to i}$ and x_k^i (see Eq. (4). Note also that as we transfer the text forward and backward, we also control the accuracy of style transfer using the classifier (see Eqs. (2), (3) and (5)). In what follows, we present the details of the loss functions employed in training.

2.1 Forward Transfer

Reconstruction Loss. Given the encoded input sentence x_k^i and the decoded sentence $\hat{x}_k^{i \to i}$, the reconstruction loss measures how well the decoder G is able to reconstruct it:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_k^i \sim X} \left[-\log p_G(x_k^i | E(x_k^i, s_i), s_i) \right]. \quad (1)$$

Classification Loss. Formulated as follows:

$$\mathcal{L}_{class\_td} = \mathbb{E}_{\hat{x}_k^{i \to j} \sim \hat{X}} \left[-\log p_C(s_j | \hat{x}_k^{i \to j}) \right]. \quad (2)$$

For the encoder-decoder this loss gives a feedback on the current generator's effectiveness on transferring sentences to a new style. For the classifier, it provides an additional training signal from generated data, enabling the classifier to be trained in a semi-supervised regime.

Classification Loss - Original Data. In order to enforce a high classification accuracy, the classifier also uses a supervised classification loss, measuring the classifier predictions on the original (supervised) instances $x_k^i \in X$:

$$\mathcal{L}_{class\_od} = \mathbb{E}_{x_k^i \sim X} \left[-\log p_C(s_i | x_k^i) \right]. \quad (3)$$

2.2 Backward Transfer

Reconstruction Loss. The *back-transfer (or cycle consistency) loss* (Zhu et al., 2017) is motivated by the difficulty of imposing constraints on the transferred sentences. Back-transfer transforms the transferred sentences $\hat{x}_k^{i \to j}$ back to the original style s_i, i.e., $\hat{x}_k^{i \to j \to i}$ and compares them to x_k^i. This also implicitly imposes the constraints on the generated sentences and improves the content preservation. The loss is formulated as follows:

$$\mathcal{L}_{back\_rec} = \mathbb{E}_{x_k^i \sim X} \left[-\log p_G(x_k^i | E(\hat{x}_k^{i \to j}, s_j), s_i) \right], \quad (4)$$

which can be thought to be similar to an auto-encoder loss in (1) but in the style domain.

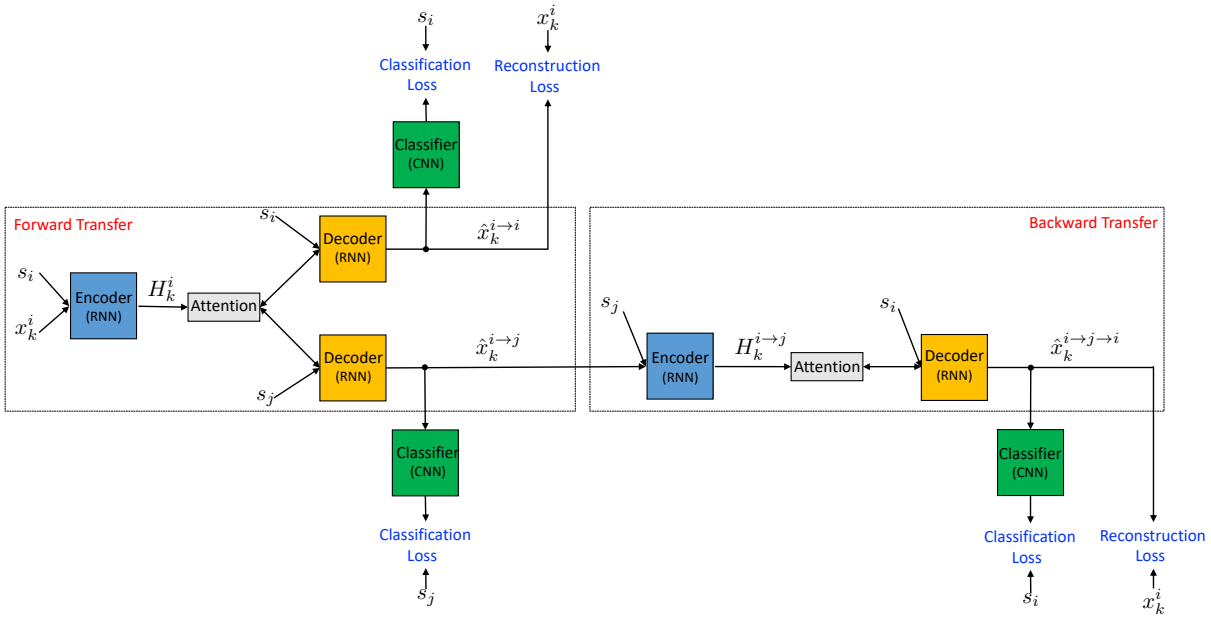

Figure 1: Proposed framework of a Neural Text Style Transfer algorithm using non-parallel data.

Classification Loss. Finally, we ensure that the back-transferred sentences $\hat{x}_k^{i \to j \to i}$ have the correct style label s_i:

$$\mathcal{L}_{class\_btd} = \mathbb{E}_{\hat{x}_k^{i \to j} \sim \hat{X}} \Big[-\log p_C(s_i | G(E(\hat{x}_k^{i \to j}, s_j), s_i)) \Big]. \quad (5)$$

In summary, the training of the components of our architecture consists in optimizing the following loss function using SGD with back-propagation:

$$\mathcal{L}(\theta_E, \theta_G, \theta_C) = \min_{E,G,C} \mathcal{L}_{rec} + \mathcal{L}_{back\_rec} + \mathcal{L}_{class\_od} + \mathcal{L}_{class\_td} + \mathcal{L}_{class\_btd}$$

3 Related Work

Most previous work that address the problem of offensive language on social media has focused on text classification using different machine learning methods (Xiang et al., 2012; Warner and Hirschberg, 2012; Kwok and Wang, 2013; Wang et al., 2014; Burnap and Williams, 2015; Nobata et al., 2016; Davidson et al., 2017; Founta et al., 2018). To the best of our knowledge, there is no previous work on approaching the offensive language problem using style transfer methods.

Different strategies for training encoder-decoders using non-parallel data have been proposed recently. Many of these methods borrow the idea of using an adversarial discriminator/classifier from the Generative Adversarial Networks (GANs) framework (Goodfellow et al., 2014) and/or use a cycle consistency loss. Zhu et al. (2017) proposed the pioneering use of the cycle consistency loss in GANs to perform image style transfer from non-parallel data. In the NLP area, some recent effort has been done on the use of non-parallel data for style/content transfer (Shen et al., 2017; Melnyk et al., 2017; Fu et al., 2018) and machine translation (Lample et al., 2018; Artetxe et al., 2018). Shen et al. (2017), Fu et al. (2018) and Lample et al. (2018) use adversarial classifiers as a way to force the decoder to transfer the encoded source sentence to a different style/language. Lample et al. (2018) and Artetxe et al. (2018) use the cycle consistency loss to enforce content preservation in the translated sentences. Our work differs from the previous mentioned work in different aspects: we propose a new relevant style transfer task that has not been previously explored; our proposed method combines a collaborative classifier with the cycle consistency loss, which gives more stable results. Note that a potential extension to a problem of multiple attributes transfer would still use a single classifier, while in (Shen et al., 2017; Fu et al., 2018) this may require as many discriminators as the number of attributes.

Another line of research connected to this work consists in the automatic text generation conditioned on stylistic attributes. (Hu et al., 2017) and (Ficler and Goldberg, 2017) are examples of this line of work which use labeled data during training.

4 Experiments

4.1 Datasets

We created datasets of offensive and non-offensive texts by leveraging Henderson et al. (2018)'s preprocessing of Twitter (Ritter et al., 2010) and Reddit Politics (Serban et al., 2017) corpora, which contain a large number of social media posts. Henderson et al. (2018) have used Twitter and Reddit datasets to evaluate the impact of offensive language and hate speech in neural dialogue systems.

We classified each entry in the two datasets using the offensive language and hate speech classifier from (Davidson et al., 2017). For Reddit, since the posts are long, we performed the classification at the sentence level. We note that since ground truth (parallel data) is not available, it is important to use the same classifier for data generation and evaluation so as to have a fair comparison and avoid inconsistencies. Therefore, we use the classifier from (Davidson et al., 2017) to test the performance of the compared algorithms in Sec. 4.3.

For our experiments, we used sentences/tweets with size between 2 and 15 words and removed repeated entries, which were frequent in Reddit. The final datasets have the following number of instances: Twitter - train [58,642 / 1,962,224] (offensive / non-ofensive), dev [7842] (offensive), test [7734]; Reddit - [224,319 / 7,096,473], dev [11,883], test [30,583] . In both training sets the number of non-offensive entries is much larger than of the offensive ones, which is not a problem since the objective is to have the best possible transfer to the non-offensive domain. We limited the vocabulary size by using words with frequency equal or larger than 70 (20) in Reddit (Twitter) dataset. All the other words are replaced by a placeholder token.

4.2 Experimental Setup

In all the presented experiments, we have used the same model parameters and the same configuration: the encoder/decoder is a single layer GRU RNN with 200 hidden neurons; the classifier is a single layer CNN with a set of filters of width 1, 2, 3 and 4, and size 128 (the same configuration as in the discriminators of (Shen et al., 2017)). Following (Shen et al., 2017), we have also used randomly initialized word embeddings of size 100, and trained the model using Adam optimizer with the mini-batch size of 64 and learning rate of 0.0005. The validation set has been used to select the best model by early stopping. Our model has a quite fast convergence rate and achieves good results within just 1 epoch for the Reddit dataset and 5 epochs for the Twitter dataset.

Our baseline is the model of Shen et al. (2017)[1] and it has been used with the default hyperparameter setting proposed by the authors. We have trained the baseline neural net for three days using a K40 GPU machine, corresponding to about 13 epochs on the Twitter dataset and 5 epochs on the Reddit dataset. The validation set has also been used to select the best model by early stopping.

4.3 Results and Discussion

Although the method proposed in this paper can be used to transfer text in both directions, we are interested in transferring in the direction of offensive to non-offensive only. Therefore, all the results reported in this section correspond to this direction.

In Table 1, we compare our method with the approach of Shen et al. (2017) using three quantitative metrics: (1) *classification accuracy* (Acc.), which we compute by applying Davidson et al. (2017)'s classifier to the transferred test sentences; (2) *content preservation* (CP), a metric recently proposed by Fu et al. (2018) which uses pre-trained word embeddings to compute the content similarity between transferred and original sentences. We use Glove embeddings of size 300 (Pennington et al., 2014); (3) *perplexity* (PPL), which is computed by a word-level LSTM language model trained using the non-offensive training sentences.

Dataset	System	Acc.	CP	PPL
Reddit	[Shen17]	87.66	0.894	**93.59**
	Ours	**99.54**	**0.933**	115.75
Twitter	[Shen17]	95.36	0.891	**90.97**
	Ours	**99.63**	**0.947**	162.75

Table 1: Classification accuracy, content preservation and perplexity for two datasets.

As can be seen from the table, our proposed method achieves high accuracy on both datasets, which means that almost 100% of the time Davidson et al. (2017)'s classifier detects that the transferred sentences are non-offensive. In terms of the content preservation, for both datasets our method also produces better results (the closer to 1 the better) when compared to (Shen et al., 2017) . A

[1] https://github.com/shentianxiao/language-style-transfer

	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

Table 2: Example of offensive sentences from Reddit and Twitter and their respective transferred versions.

what *big* century are you living in ?
life is so *big* cheap to some people .
you 're *big* pathetic .

Table 3: Examples of common mistakes made by our proposed model.

reason for these good results can be found by checking the examples presented in Table 2. The use of the back transfer loss and the attention mechanism makes our model good at preserving the original sentence content while being precise at replacing offensive words by the non-offensive ones. Also observe from Table 2 that, quite often, Shen et al. (2017)'s model changes many words in the original sentence, significantly modifying the content.

On the other hand, our model produces worse results in terms of perplexity values. We believe this can be due to one type of mistake that is frequent among the transferred sentences and that is illustrated in Table 3. The model uses the same non-offensive word (e.g. *big*) to replace an offensive word (e.g. *f***ing*) almost everywhere, which produces many unusual and unexpected sentences.

We have performed ablation experiments by removing some components of the proposed model. The results for the Twitter dataset are shown in Table 4. We can see that attention and back-transfer loss play important roles in the model. In particular, when both of them are removed (last row in Table 4), although the classification accuracy improves, the perplexity and the content preservation drop significantly. This behavior happens due to the trade off that the decoder has to balance when transferring a sentence from a style to another. The decoder must maintain a proper balance between transferring to the correct style and generating sentences of good quality. Each of these properties can easily be achieved on its own, e.g., copying the entire input sentence will give low perplexity and good content preservation but low accuracy, on the other hand, outputting a single keyword can give high accuracy but high perplexity and low content preservation. While the classification loss guides the decoder to generate sentences that belong to the target style, the back transfer loss and the attention mechanism encourage the decoder to copy words from the input sentence. When both back transfer loss and attention are removed, the model is encouraged to just meet the classification requirement in the transfer step.

System	Acc.	CP	PPL
Full	99.63	**0.947**	**162.75**
No Attention	99.88	0.939	196.65
No Back Transfer	97.08	0.938	257.93
No Att & Back Trans	**100.0**	0.876	751.56

Table 4: Ablation results for the Twitter dataset.

It is important to note that current unsupervised text style transfer approaches can only handle well cases where the offensive language problem is lexical (such as the examples shown in Table 2), and just changing/removing few words can solve the problem. The models experimented in this work will not be effective in cases of implicit bias where ordinarily inoffensive words are used offensively.

5 Conclusions

This work is a first step in the direction of a new promising approach for fighting abusive posts on social media. Although we focus on offensive language, we believe that further improvements on the proposed methods will allow us to cope with other types of abusive behaviors.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural ma-

chine translation. In *International Conference on Learning Representations*.

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

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*. pages 512–515.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Workshop on Stylistic Variation*.

A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. 2018. A unified deep learning architecture for abuse detection. *ArXiv e-prints* .

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of AAAI*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. pages 2672–2680.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the AAAI/ACM conference on Artificial Intelligence, Ethics, and Society*.

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*. pages 1587–1596.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. pages 1621–1622.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Igor Melnyk, Cicero Nogueira dos Santos, Kahini Wadhawan, Inkit Padhi, and Abhishek Kumar. 2017. Improved neural text attribute transfer with non-parallel data. In *NIPS Workshop on Learning Disentangled Representations: from Perception to Control*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. pages 145–153.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *The 2010 Annual Conference of the NAACL*. pages 172–180.

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

T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. pages 415–425.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. pages 19–26.

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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* .