Offensive Language Translation and Hate Speech Detection

Grace Lee, Anstonia Ma, Mon Young
University of California, Berkeley
School of Information, Master's in Data Science
{grace_lee, anstoniama, mon.young}@berkeley.edu

Abstract

We present an approach to clean comments on social media with offensive language translation and discuss how this relates to hate speech detection. We first fine-tune pre-trained translation and summarization models on a paralleled offensive-language and detoxified dataset. Next, we train a multi-class classifier on tweet data containing hate speech, offensive, and neutral language labels. We then assess the translation models on the tweet data using standard NLP translation quality metrics and manual evaluation. We find that for a same-language offensive language "translation" task, summarization-based models may perform the best, and that our translation models actually improve hate speech classification.

1 Introduction

Hate speech and offensive language has a persistent presence on the internet, especially on social media networks, and there is constant discussion and work on reducing such content from these forums. Social media companies such as Facebook and Twitter have faced criticism for allowing hate speech to perpetuate on their platforms and content moderators are frequently employed by organizations to filter offensive posts. A simple way to "cleanse" forums would be to censor and remove all comments with offensive language, but we believe there may be relevant content in some of these posts that would be useful to keep. Comments whose underlying meaning is harmless but contain offensive language should be differentiated from comments containing harmful content such as hate speech.

In this work, we aim to create a translator that detoxifies content with offensive language while retaining the original meaning and identify what effect translating offensive language has on identifying instances of hate speech. We train a variety of style transfer models for offensive language translation and classification models to identify whether comments are neutral, hate speech, or contain offensive language. We then evaluate the usefulness of our translation models by comparing the classification results of original and "detoxified" translations.

2 Background

There has previously been research to filter out hate speech and offensive posts (Xiang et al., 2012, Wang et al., 2014; Nobata et al., 2016; Davidson et al., 2017). These works use various methods for text classification to identify offensive comments and operate under the assumption that posts with offensive language should be censored entirely. The research by Davidson et al. (2017) places particular emphasis on the distinction between hate speech and general instances of offensive language. These concepts are related, but not completely identical. Using their definition, hate speech is "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. . . and does not include all instances of offensive language." This distinction is important because general classification methods might cause posts to be labeled as hate speech simply with the presence of offensive language, causing high false positive rates and low precision.

There is also additional research (Nogueira dos Santos et al., 2018; Logacheva et al., 2022) to translate offensive language to non-offensive language, offering an alternative to censoring the entire post. Both used encoder-decoder style-transfer approaches to rewrite posts while maintaining content, but had different methods depending on the presence of a "parallel" offensive and non-offensive dataset. Nogueira dos Santos et al. (2018) did not have parallel data available, so they adjusted their encoder-decoder models with additional collaborative classifiers to indicate the effectiveness of the decoder portion of their model. In contrast, the

research by Logacheva et al. (2022) suggested a new, quick pipeline to collect paralleled data via crowdsourcing rather than using experts. With this paralleled dataset, they fine-tuned pre-trained language models such as BART for the detoxification task.

We expand on the offensive language translation task with the Logacheva et al. (2022) parallel data by training additional models. We also look into classifying hate speech vs. offensive language, and how translation models impacts hate speech detection.

3 Methods

3.1 Datasets

For our translation models, we used the ParaDetox dataset[1] created by Logacheva et al. (2022), which contains 19,766 pairs of sentences; one of the sentence pairs is toxic, and the other is neutral. This sentence pair creation was crowdsourced; non-toxic labels were manually translated and verified in their crowdsourcing pipeline to generate a neutral version of the sentences with the same meaning. Using this parallel dataset, we are able to use pre-trained style-transfer methods without needing to add additional classifiers.

For our classification models, we used the dataset from the Davidson et al. (2017) research[2], which contains tweets from Twitter with three category labels: hate speech, contains offensive language (but not hate speech), and neutral.

3.2 Setup

We fine-tune offensive language translation models on the parallel ParaDetox dataset to translate offensive language. We also train a classification model, labeling posts as hate speech, offensive language, or neutral, on the Tweet dataset. We then run our translation models on the tweets to evaluate the effectiveness of our translations and impact on hate speech on "unseen" data.

A diagram of our process can be found in Appendix A.

3.3 Models

Our first task is the offensive language translation. Our baseline model is the ParaDetox BART model created by Logacheva et al. (2022). We also train zero-shot BART Large, Bart XSUM, Bart CNN, and T5 Summarization models. Because this translation is in English for both the original and translation, the model tends to keep a majority of the original words. This is different from other word-to-word translations such as English to Chinese or English to German, and could potentially be seen as a form of summarization. We are interested in the prediction differences between translation and summarization models. We chose BART Large for its general-purpose translation, BART CNN for its domain-specific translation, and BART XSUM and T5 for summarization. We expect most posts in comment sections and social media to be relatively short, so we fine-tune the models with maximum sequence length of 25.

Our second task is the hate speech and offensive language classification. For this task, we fine-tune BERT-based models on the Davidson et al. (2017) Twitter data. Inspired by the Nogueira dos Santos et al. (2018) paper, we created a classification baseline model utilizing a word2vec CNN and fine-tuned three additional BERT models: BERT Uncased, BERT Large, and BERTweet.

For the classification task, stratifying our data preserved the class distribution throughout our training, validation, and test sets. However, preliminary EDA revealed that there was a much larger number of offensive language tweets than hate speech and neutral labeled ones. To combat this, we utilized the class_weight parameter and determined it using the following formula:

$$\frac{Total\ samples}{(\#\ of\ classes) * (samples\ per\ class)}$$

Interestingly enough, the additional class_weights penalty decreased our model's general accuracy as the accuracy for the majority (offensive language) class declined while accuracy for the other two (hate speech and neither) increased. As we want to have accurate classification of all our classes, we proceeded with this class_weight parameter despite the decrease in general accuracy.

3.4 Evaluation

We determine the "best" BERT classification model using accuracy as determined by the number of predicted classes that match the true labeled class. Since we are interested in hate speech, we also

[1]ParaDetox dataset can be found here: https://huggingface.co/datasets/SkolkovoInstitute/paradetox

[2]Davidson tweet dataset can be found here: https://paperswithcode.com/dataset/hate-speech-and-offensive-language

focus on the accuracy of the hate speech class in addition to the model's overall accuracy.

For the translation models, we use several evaluation methods. With the ParaDetox test set, we evaluate the "detoxified" translations compared to the parallel original comments with BLEU, BLEURT, and METEOR. We selected BLEU for its simplicity and a basic evaluation of translation quality, and selected BLEURT and METEOR as different metrics that correlate better with human judgment. We also looked at ROUGE scores in consideration of the BART XSUM and T5 models. We also looked at semantic similarity by looking at the cosine similarity of the original and translated sentences put through Universal Sentence Encoder and also the Semantic Textual Similarity benchmark[3].

Since the translation models were trained on the ParaDetox dataset, we also wanted to evaluate the models on a new, unseen dataset. We ran our four translation models on the Davidson et al. (2017) tweet data and manually evaluated the quality of those translations. We took a sample of 73 tweets (original and translated pairs) and gave them a binary rating on the following questions:

- Did translation remove offensive language?
- Was the meaning of the original tweet changed?
- Did translation cut off a large part of the tweet?
- Fluency: how human readable are the translations?

We also wanted to answer the question of "how do the translation models impact hate speech detection?" After translating the test set of the Davidson et al. (2017) Tweet data, we applied our best BERT classification model to the original and translated tweet pairs to get a like-for-like comparison of the change in hate speech classification. We did this four times for the following translation models: ParaDetox, BART CNN, BART XSUM, and T5. For this segment, we consider the combination of our best classification model and the ParaDetox translation model to be the baseline.

4 Results

4.1 Hate Speech Classification

Model accuracy for all four models were fairly similar, with the exception of the BERT Large model. We selected the fine-tuned BERTweet model to be our final classification model. Although its general accuracy was the lowest of the three remaining models, the class-specific accuracy for hate speech was the highest. Since we are most interested in the hate speech classification, we decided to consider the hate speech accuracy to be important. Both the BERT Uncased and BERTweet models were better than the baseline for this metric, with BERTweet having the highest accuracy at 86.7% in this category. We also felt the BERTweet model maintained similar accuracy for the other classes. In addition, the pre-trained BERTweet model was trained on tweets, which is a similar lexicon to the Davidson et al. (2017) tweet dataset, which will be used in the final evaluation of the translation models.

Model	Loss	Acc.
CNN Baseline	0.729	0.645
BERTweet	0.703	0.687
BERT Uncase	0.765	0.558
BERT Large	0.058	1.132

Table 1: Model Loss and General Accuracy

Model	Hate Accuracy	Offensive Accuracy	Neither Accuracy
CNN Baseline	0.434	0.737	0.789
BERTweet	0.867	0.656	0.866
BERT Uncase	0.762	0.738	0.887
BERT Large	1.000	0.000	0.000

Table 2: Class Accuracy

4.2 Offensive Language Translation

When building the translation models, we fine-tuned the models based on improving BLEU, BLEURT, and METEOR metrics. As we began exploring other methods for translating offensive language, we started using models typically used for summarization, such as BART XSUM and T5. Once we added these models, we also started considering ROUGE metrics during the fine-tuning process.

In this first evaluation stage, we noticed that the baseline Logacheva et al. (2022) ParaDetox model performed the best in all metrics when evaluating on the ParaDetox dataset. However, we expect this is due to Logacheva et al.'s extensive work training the model on the matching ParaDetox dataset, and expect this might change on new data.

[3]https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

Model	Meteor	BLEU	BLEU RT	ROUGE
BART Large	0.805	0.479	0.218	0.813
BART XSUM	0.803	0.494	0.213	0.817
BART CNN	0.822	0.475	0.222	0.823
T5	0.814	0.477	0.203	0.818
ParaDetox	0.844	0.597	0.366	0.860

Table 3: ParaDetox Dataset Translation Evaluation

Model	Meteor	BLEU	BLEU RT	USE	STS	ROUGE
T5	0.717	0.644	-0.124	0.763	0.812	0.828
XSUM	0.646	0.540	0.000	0.780	0.776	0.746
CNN	0.705	0.602	0.094	0.803	0.795	0.792
ParaDetox	0.715	0.605	0.674	0.807	0.803	0.799

Table 4: Davidson Tweet Dataset Translation Evaluation

Of the four models we trained, BART XSUM had the best BLEU score and BART CNN had the highest BLEURT and METEOR scores as can be seen in Table 3. Interestingly, the BART CNN model also had the highest ROUGE score despite it not traditionally being a summarization model. As we moved on to the next stage of evaluation, we decided to drop the BART Large model since the other BART models were better, and kept the T5 model as we were interested to see how a different model would work with unseen data.

4.3 Translations on Unseen Tweet Data

In our next phase evaluating the translation models, we wanted to check the translation quality when using our models on unseen data. We ran the four models, the baseline ParaDetox, BART XSUM, BART CNN, and T5 on the Davidson et al. (2017) tweet data.

4.3.1 Translation Quality

We again look at BLEU, BLEURT, METEOR, and ROUGE to evaluate translation quality on the tweet data, shown in table 4. With these metrics, the T5 model appears to be the best translation on this new data as it had the highest METEOR and BLEU scores. From a summarization perspective, T5 was also the best model as it had the highest ROUGE score. The T5 model was also our only model that outperformed the baseline ParaDetox model on these metrics.

Since this data has no parallel "detoxified" version, we also were interested in semantic similarity as another way to compare the original and translated tweets. For this task, we looked at the cosine similarity after using Universal Sentence Encoder (USE) and the Semantic Textual Similarity benchmark (STS). The USE cosine similarity ranges from 0 to 1 and the STS benchmark is normalized from -1 to 1. All models had fairly high USE and STS scores, indicating a high degree of similarity between the original and translated tweets. This was a good sign for us as these scores suggest our model translations are maintaining the original meaning. The T5 model had the highest similarity from an STS perspective. The baseline ParaDetox model, followed closely by the BART CNN model, had the highest USE similarity.

On this new data, the ParaDetox baseline is no longer the best in a majority of these metrics. This confirms our assumption that the ParaDetox model works better in its own dataset, while our models have better performance on unseen data.

4.3.2 A Brief Aside on BLEURT

We noticed the BLEURT scores were frequently negative (indicating the translation was either changing the meaning or not fluent) while training the translation models. The final T5 model, which we considered the best in most other aspects, continued to have negative BLEURT scores. We investigated instances where the T5 translations had the worst BLEURT scores:

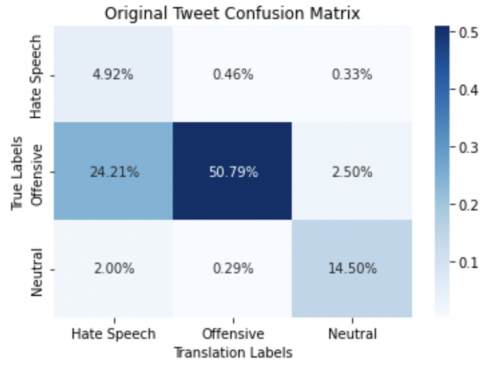
Original	Translation
n*gger mad	I'm mad
Fat white b*tch ad a skinny po' black n*gga..... yeah im in Metairie f sho...	I'm in Metairie, Louisiana

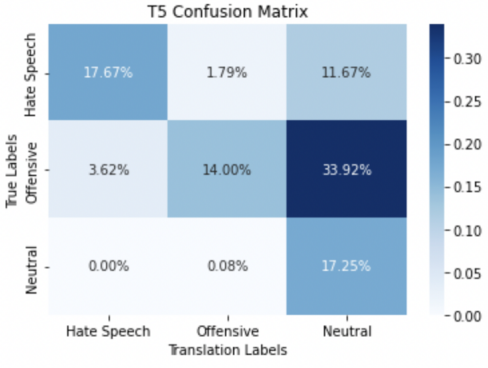
Table 5: T5 Translated Comments

Overall, we believe the translations look fine and are generally human readable. There were instances where the original tweet's language was not fluent, causing the resulting translation to also look slightly strange from a fluency perspective. Although BLEURT is typically used as a translation quality metric that takes into account human readability, it is perhaps not a great metric for the tasks in this paper.

4.3.3 Classification Comparison

We noticed that all translation models improved the hate speech classification. In the BERTweet classification of the original tweets, 4.92% of the sample were accurately labeled as hate speech, with 24.21% offensive language being labeled as hate speech and 2% of neutral language being labeled as hate speech. The BERTweet classification on the translations of all the other models had higher percentages of the sample accurately classifying hate speech, with T5 notably having the highest at 17.67%. For the translation models, the percent of the sample that was offensive language or neutral but incorrectly classified as hate speech was also reduced to under 5%. We show the confusion matrix for the original and T5 translated tweets here, confusion matrix for all other models can be found in Appendix B.





4.3.4 Manual Evaluation

We took a sample of 75 offensive and detoxified tweet pairs and manually gave them a binary rating across four categories: Removed Offensive Language (ROL), Changed Meaning of Original (CM), Cut Out Large Portion of Tweet (COT), and Fluent and Human Readable (FHR)[4]. These findings are summarized in Table 6:

Model	ROL	CM	COT	FHR
T5	64.5%	2.8%	1.4%	87.5%
XSUM	75.8%	3.8%	15.4%	100%
CNN	76.7%	5.5%	13.7%	95.9%
ParaDetox	70%	9.9%	5.6%	98.6%

Table 6: Davidson Tweet Dataset Translation Evaluation

The T5 model performed the best when we looked at the metrics for translation quality, and our manual evaluation confirms this as the T5 model had the lowest percentage of cases (2.8%) where the translation changed the meaning of the original tweet. However, there were other limitations of the T5 model. The T5 model did the worst at removing offensive language and was the least human readable, performing even worse than the baseline ParaDetox model. On these fronts, the BART-based XSUM and CNN models perform better. However, these two models also cut out large portions of the original tweets.

It is difficult to conclude which is the "best" model from our manual evaluation, which may be due to the small sample size. The T5 model seemed to stay the most "true" to the original tweets at the cost of readability and detoxification. On the other hand, XSUM and CNN were better for the end-user to read, but at the cost of potentially losing information.

Across all models, the translations were generally successful and the meaning of the original tweets were largely preserved. However, there were some patterns where all models repeatedly missed certain offensive language. For example, we observed that "hoe" was almost never translated while a similarly offensive word "b*tch" would almost always be changed to "girl" or deleted altogether.

[4]We defined "cut out" to be where meaningful parts of the tweet was cut out. This might not impact the other metric "changed meaning" if the non-cut out portion did not change the meaning. "Fluent and human readable" should also be caveated: if the original tweet was not very fluent, we decided the translation to be fluent if it closely aligned with the syntax of the original

We suspect that this may be due to the lexicon differences between the datasets as the Tweet dataset classifies "hoe" as an offensive word while the ParaDetox set may not have as many instances of it. Another theory is that because hoe is also a descriptive word for a gardening tool, the usage in the tweets is not seen as offensive.

4.3.5 A Deeper Look at T5 and Hate Speech

As we judged the T5 model to be the best model from the translation quality scores and improvement in hate speech classification, we took a deeper look at the translations from this model. In all instances where the BERTweet model classified the original as hate speech and the translation as neutral, as well as where the model classified the original as offensive language and the translation as neutral, T5 was able to accurately remove the offending words with some fluency. However, there were some interesting cases of previously classified as offensive language becoming classified as hate speech, even though the offending language was removed entirely.

Original offensive language comment	Translated comment (categorized as hate speech)
I can't stand a stuck up b*tch. Always got a attitude for no reason. I don't know who hurt y'all but stay TF from round me ;;;	I can't stand a stuck up person. Always got an attitude for no reason. I don't know who hurt you but stay away from me.
Why would you wanna be the Green Ranger? He's evil! Duh b*tch	Why would you wanna be the Green Ranger? He's evil

Table 7: Davidson Tweet Dataset Translation Evaluation

We see above that the two translations do not contain hate speech, but do have some other language that may be used in hateful comments such as "evil" and "stuck up" or "hurt." Additionally, there does seem to be a theme that our model is better at removing offensive language of the cuss word variety compared to hate speech. We see this in the following example:

Original hate speech comment	Translated comment (categorized as offensive language)
What a f*cking f**got	What a f*ggot! ! ! ! ! ! !

Table 8: Davidson Tweet Dataset Translation Evaluation

5 Conclusions

We created several translators that detoxifies offensive language while largely retaining the original comment's meaning. We also evaluated the translators and looked into each model's impact on hate speech classification, using a fine-tuned BERTweet model as the classifier. The baseline ParaDetox model appeared to be the best translation model when testing on the ParaDetox data, but this does not hold up on unseen data. When translations were run on the Tweet data, it looked like T5 was the best model in terms of fluency and retaining meaning from the translation quality and semantic similarity metrics. Offensive language translation also improved hate speech classification, with T5 having the most improvement. The sample size from our manual evaluation made it difficult to make conclusive insights, but we did find some strengths and weaknesses of the models and also found interesting patterns in the translations. For same-language (i.e. English to English) translation of offensive language, summarization methods may be better than translation methods, as can be seen by the T5 model being the best.

With some initial success in offensive language translation, we believe there is further opportunity to improve the translation models. Perhaps with another iteration of translation, comments currently still "offensive" may become inoffensive. Another improvement would be to continually add to our parallel corpus through a self-supervising process to train and improve our style transfer models. Adding additional datasets can provide additional lexicon that do not exist in the ParaDetox data to further build out our model. Finally, even with differentiating hate speech from offensive language, there are other sub-categories within these broader umbrellas that would be worth classifying, such as racism, sexism, ableism, and the like which each have their own nuanced appearances and usage.

References

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

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Pages 6804 – 6818.

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

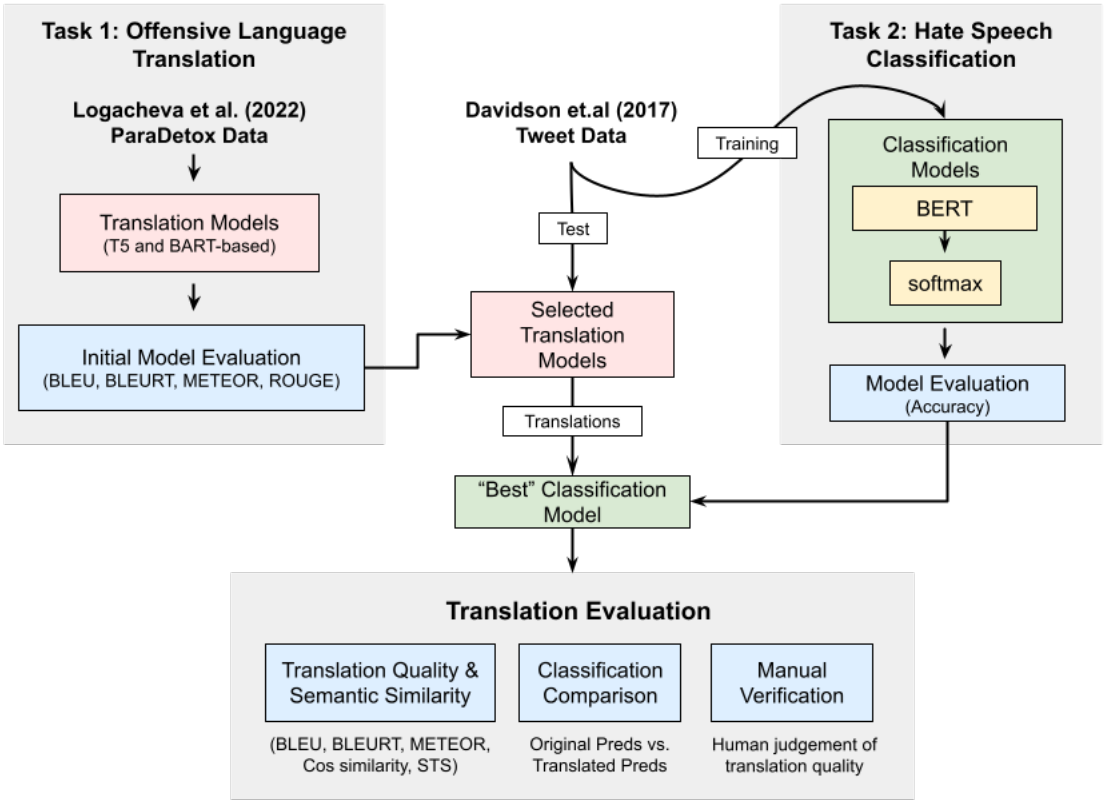
Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Pages 189 – 194.

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

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. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. Pages 14980 - 1984.

Appendix

A Diagram of Process



B Confusion Matrices