Men Have Feelings Too: Debiasing Sentiment Analyzers using Sequence Generative Adversarial Networks

Athiya Deviyani[1], **Haris Widjaja**[1], **Mehak Malik**[1], **Dan Hoskins**[1]

[1] Carnegie Mellon University

adeviyan@cs.cmu.edu, iwidjaja@cs.cmu.edu, mehakm@cs.cmu.edu, dhoskins@cs.cmu.edu

Abstract

Natural Language Processing (NLP) models often magnify biases with respect to race, gender and age present in datasets that they are trained on. Furthermore, it is becoming increasingly challenging to collect an unbiased dataset given that sexist and racist content are ubiquitous in common sources of data such as social media. In this work, we propose a Generative Adversarial Network-based approach to augment a sentiment analysis dataset to mitigate the gender biases present in the original dataset. Ultimately, by evaluating on a downstream sentiment analysis task using a model trained on the augmented dataset, we show that our method successfully reduces the disparity between the sentiment scores across the different genders, while maintaining the overall model performance.

Introduction

As the performance of Natural Language Processing (NLP) systems rapidly improve across many subdomains (e.g. translation, sentiment analysis, topic classification), they earn more trust among the general public. It's easy to see why: when a system has a very high accuracy, it seems to be functioning very well. As a result of this perception, more of the general public will believe the benefits outweigh the risks of these systems, and their adoption increases. While these systems are unquestionably useful in many scenarios, a number of pernicious effects of these NLP-based systems exist, including biased predictions. These biased predictions result in unfair outcomes for marginalized groups. As the adoption of NLP models increases, the deleterious effects of these biased predictions will commensurately be propagated.

For example, human resource departments in private corporations are adopting sentiment analysis tools for gauging employee feedback. These systems are used to better inform decisions about the future directions of the companies that use them. Unfortunately, it has been found that numerous sentiment analysis systems exhibit gender bias. One review by Maurer (2019) found that, of all classifiers that took part in the "Affect In Tweets" SemEval task (Mohammad et al. 2018), most systems output higher sentiment intensity predictions for one gender than another. Since the sentiment associated with any given piece of feedback can determine the extent to which it is considered in decisions, these biased predictions can result in one gender's feedback being considered more than the other. This will, in turn, impact company decisions, which affects employees. Clearly, NLP-based systems' biases can impact decisions, resulting in detrimental downstream effects on humans. Therefore, it is critical to develop methods to combat these biases, enabling the machine learning community to build fairer systems.

Of the existing bias-mitigation techniques, many focus on modifying the dataset that the systems are trained on. These techniques have serious limitations. For example, counterfactual data augmentation (i.e. adding sentences with swapped pronouns of existing data points to the dataset) requires manual labeling. This must be performed for every new task that the technique is applied to. This means that developers of these systems have to dedicate significant effort to modify their datasets to mitigate the biases.

It would be valuable to have a bias-mitigation technique that generalizes well to arbitrary contexts, eliminating the need for significant rework. One technique uses generative adversarial learning to generate data points with which to augment the training dataset. In this technique, the relationships necessary for bias mitigation are learned, rather than specified through manual labels, suggesting it might generalize to arbitrary contexts. However, this generalizability hasn't been thoroughly tested.

Problem definition

The high-level goal of our work is to obtain a generalizable method to neutralize bias in various NLP datasets through data augmentation using the adversarial learning objective. We believe that training a model on the resulting augmented debiased dataset will lead to fairer results as defined by the reduction of scoring or performance disparity.

The reason we

We will evaluate the resulting dataset from our methodology using sentiment analysis as the downstream task. Given a text sequence, a sentiment analyzer model will decide whether the the text sequence expresses negative or positive sentiment. This is essentially a binary text classification task. Following from the observation made by Kiritchenko and Mohammad (2018), we hypothesize that sentences con-

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

taining female-related terms (such as “mother”, “woman”, “girl”), amplifies the sentiment of the sentence. For example, the sentence “my sister is sad” will have a higher negative sentiment score when compared to the sentence “my brother is sad”, given that the former contains a female-related term (“sister”).

We believe that the amplification of sentiment can have adverse effects on people identifying as female. In addition to enforcing dangerous stereotypes, there are many cases where sentiment analysis is used to automatically detect depression from social media or blog post entries to advance psychological research and mental health aid, similar to what Husseini Orabi et al. (2018) and Deshpande and Rao (2017) has done in their work. The misclassification of sentiment for women (and potentially other minority groups) may lead to problems such as having one gender group being flagged as more depressed than the other.

NLP applications are also commonplace in professional settings. As mentioned previously, some human resource departments within organizations increasingly use automated sentiment analyzers to assess employee feedback (Maurer 2019). This entails that biased sentiment analysis results can lead to the silencing of minority voices, wherein their feedback might not be taken into account. Additionally, the over-amplification of sentiment on text sequences containing instances of female-related words can also mean that negative words are scored more severely, therefore a female worker who received negative feedback might be scored lower than a male worker who received a similar kind of feedback. Conversely, if positive words are also scored more highly, then it would be unfair for male workers who receive similar positive feedback but may receive a much lower positive score.

Therefore, our main objective in this project is to use the adversarial training objective to artificially generate text that simulates real text sequences with the aim of neutralizing the effect of data disparity to the over-amplification of sentiment in the sentiment analysis task on text containing female-related terms. Given that our method works in mitigating biases of the downstream task, our method can potentially be adapted to various tasks and validates the generalizability of a novel, reproducible debiasing technique.

Dataset

We decided to use the Sentiment140 dataset as our baseline collected by Go, Bhayani, and Huang (2009). The dataset contains over 1.6 million tweets, alongside other information such as the tweetID, tweet date, query used to obtain the tweet, and the tweet author. Each tweet comes with the corresponding sentiment/polarity label of NEGATIVE and POSITIVE. Below are examples of tweets reflecting the aforementioned sentiment labels:

Contrary to most datasets collected for the task of sentiment analysis, the Sentiment140 dataset’s labels are not hand-annotated. In fact, they have automated the collection process fully. The authors collected a handful of tweets from twitter and automatically labeled tweets containing positive emoticons such as `:)` as POSITIVE and tweets containing emoticons that signify negative emotions such as `:(` as NEGATIVE. We have performed Twitter-specific preprocessing on top of normal text processing (which involves lowercasing, tokenizing, and removing stop words), such as obfuscating and removing usernames and links for privacy purposes, as well as separating hashtags into individual words (e.g. #BlackLivesMatter → Black Lives Matter) using the help of the `ekphrasis` Python library (Baziotis, Pelekis, and Doulkeridis 2017). This step is highly crucial as we would like to minimize the amount of noise in our dataset (such as the presence of stop words or non-English words) while retaining as much information with respect to the sentiment label as possible. Since our task is to mitigate biases with respect to sentiment classifiers between the two gender classes, we were careful to not remove stop words which are male or female-related, such as “he”, “she”, “hers”, “his”, and more.

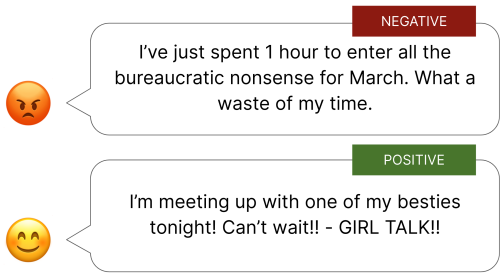

Figure 1: Example depicting NEGATIVE sentiment (up) and POSITIVE sentiment (down)

After cleaning the tweets, we decided to perform an initial visual exploration on the dataset to further understand the data. We have used the `WordCloud` Python library to create visualizations which show which words occur most frequently in tweets labeled as containing a NEGATIVE sentiment and tweets labeled as containing a POSITIVE polarity. From Figure 2, we can see that there exists words such as “bad”, “hate”, and “miss” in tweets classified as NEGATIVE while the POSITIVE-labeled tweets contain words such as “love”, “like”, and “haha”. We are aware that this signifies the tendency of a model to learn the correlation between the existence of such words and the final predicted sentiment label. During model training, we have split the dataset into 60% training data, 20% validation data, and 20% test data.

Metrics

Since our main objective is to reduce bias earlier in the pipeline by augmenting the dataset through artificial text generation, we are more interested in the fairness of the model rather than the overall sentiment classification performance of our model. However, we are also keeping track of the overall classification performance using accuracy and F1-score, as well as the precision and recall for each class with respect to our downstream sentiment analysis task, since we also want to prove that training on the augmented dataset still results in a model with a performance that is good enough to be useful.

To quantitatively evaluate model fairness, we will calculate the average sentiment score for each class, as well as the minimum and maximum sentiment score. This is

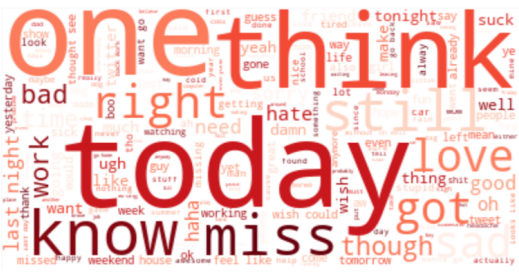

Figure 2: WordClouds depicting the most common words found in tweets labeled as NEGATIVE (up) and POSITIVE (down)

done by identifying the gender-related terms (such as "mother"/"father", "actor"/"actress", etc.) within a tweet and the corresponding sentiment prediction. A predicted score of greater than or equal to 0.5 signifies a positive sentiment while a score of less than 0.5 signifies a negative sentiment. Our main objective would be to try and minimize the difference between the average sentiment score for text containing male and female terms.

We use this metric as mentioned in Sun et al. (2019) because we assume in a large corpus that the number of statements of positive and negative sentiment is roughly the same, i.e. there is no skew with respect to gender. A better metric however may be to measure the difference in sentiment scores assigned to statements which only differ in the gender which they address. For example measuring the difference in score between statements such as "He is unhappy" versus "She is unhappy" might be more appropriate and this is an evaluation method we wish to implement in the future.

Finally, we will compare the performance of the model trained on the original Sentiment140 dataset and the performance of the model on the new augmented dataset in the sentiment analysis task. We will keep the model identical for a fair comparison without additional dataset-specific hyperparameter tuning. We will calculate the performance of the model on a held out test dataset which is a subset of the Sentiment140 dataset and contains original ground truth labels. We used the metric as Each metric will be calculated independently for text containing male and female-related words, but we will also present the overall performance metrics for each model for a general comparison.

Related work

Dataset debiasing using generative methods

This work takes inspiration from Agrawal (2022), which proposes to debias a dataset for a natural language task by augmenting it with synthetic samples generated by a language model generator. The debiasing is achieved by training the generator to avoid generating sequences with strong indication of certain protected attributes. Agrawal (2022) attempts this framework on the task of identifying conversational tweets from non-conversational tweets and using author ethnicity as a protected attribute. We build on Agrawal (2022) by attempting their proposed method to a sentiment analysis task and by using gender as protected attribute, demonstrating the generalizability of this framework.

Generative Adversarial Networks (GANs) for natural language generation

To train the generator to avoid generating sequences with strong indication of ethnicity, Agrawal (2022) uses an ethnicity prediction model as a feedback mechanism in a GAN-like framework. One challenge in adopting the original GAN framework (Goodfellow et al. 2014) to natural language generation is the inability to backpropagate through the sampling operation which language models use to sample words from logits.

To address this difficulty, Agrawal (2022) adopted a reinforcement learning-based approach, SeqGAN (Yu et al. 2017), which bypasses the need to backpropagate through the sampling operation. Motivated by shortcomings of maximum likelihood training (Welleck et al. 2019), the original SeqGAN uses a real-fake discriminator to train a generator to generate realistic language. Since this allows for an arbitrary form and number of discriminators, this makes SeqGAN an attractive framework for controlling the characteristics of generated text.

In this work, we aim to investigate the SeqGAN framework's ability to train generators to generate data useful for specific downstream tasks, while ensuring that the generated samples adhere to a non-sexism constraint, leading to an overall less biased augmented dataset for downstream models.

Baseline sentiment analysis model

Text classification model

To determine the effectiveness of our proposed debiasing approach, we train a deep learning model on the sentiment analysis task with no further intervention. Our baseline model consists of two bidirectional Gated Recurrent Unit (GRU) layers followed by a convolutional layer, a Recurrent CNN architecture shown to be effective for text classification by Zhang et al. (2018). The GRU preserves historical information in long text sequences, while the final convolution layer extracts local features, leading to a better representation for the sentiment classification task. We use a hidden size of 64 for both the GRU cells and the convolution layer, with a dropout layer after each bidirectional GRU layer to regularize our model and prevent overfitting. The network architecture is shown in Figure 3.

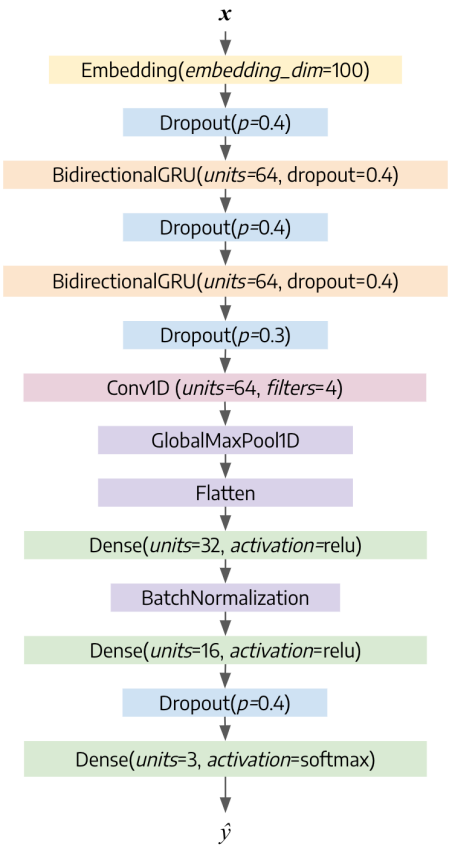

Figure 3: Recurrent CNN architecture for sentiment classification

The model takes in special GloVe word embeddings with 100 dimensions which were trained on tweets as input (Pennington, Socher, and Manning 2014). We believe that this is the most suitable representation for our training data as it consists of highly colloquial text from tweets, and that this feature representation will capture the relationships between the words most appropriately, retaining as much of its original meaning as possible in a vectorized form.

Results and analysis

We measured our baseline results using the overall performance and sentiment scores as described in Section 2.3. The classification performance is tabulated in Table 1.

	Precision	Recall	F1-score	Support
Negative	0.81	0.83	0.82	3625
Positive	0.73	0.69	0.71	2375

Table 1: Baseline sentiment analysis performance

The baseline model performs with an average accuracy of 78% on the test dataset. It is also interesting to see that the model performs better in predicting negative sentiment than positive sentiment. The F-1 score for predicting negative sentiment is 0.82 and 0.71 for positive sentiment.

		Mean	Minimum	Maximum
Negative	Male	0.27	0.03	0.92
	Female	0.08	0.01	0.82
Positive	Male	0.73	0.08	0.96
	Female	0.92	0.18	0.99

Table 2: Sentiment scores for the baseline model

Upon examining the sentiment scores of our baseline in Table 2, we see that there is a significant difference between the mean sentiment scores for different genders. We can see that for statements containing nouns or pronouns associated with females, the mean positive sentiment (0.92) is much higher than the mean positive sentiment associated with males (0.73). Similarly, the mean negative sentiment is much lower for the female class (0.08) compared to the mean negative sentiment of the male class (0.27). We can also see that sentiment scores for females are more extreme as observed by the minimum and maximum scores shown in Table 2. The mean positive sentiment score being higher and the mean negative score being lower for females can be attributed to the stereotype that women are more emotional than men and it is possible that the dataset contains more data for women that is exaggerated in terms of sentiment. This is the disparity that our work aim to mitigate by generating a more balanced dataset through augmentation.

Methodology

Our approach focuses on using the SeqGAN framework to generate an unbiased dataset which can be used to augment a biased dataset to be used in a downstream NLP task. The downstream NLP task to be used here is sentiment analysis and the objective will be to generate a gender-agnostic dataset by augmenting the Sentiment140 dataset.

Overview

The overall architecture of our approach is described in Figure 4. The approach involves using a Generative Adversarial Network (GAN) to generate synthetic tweets that are similar to our original dataset, the Sentiment140 dataset. We implement the SeqGAN as described by Yu et al. (2017) which will be explained in Section 5.2. The generator used in the GAN is DistilGPT2 model trained on tweets (Dayma 2021). The generated synthetic tweets are then subject to a selection process by a sexism detector. For this, we used the classifier described by Safi Samghabadi et al. (2020), which will be explained further in Section 5.3. We use this classifier to remove tweets that contain sexist nuances which might increase the gender bias in our dataset. Finally, we use a sentiment analyzer from Heitmann et al. (2020) as our 'gold standard' sentiment analysis tool to label our generated data. The process of generating data using the SeqGAN and purging the gender biased data to produce unbiased data can be repeated many times to increase the size of our augmented dataset. We then train a model identical to our baseline using the augmented dataset and evaluate the results.

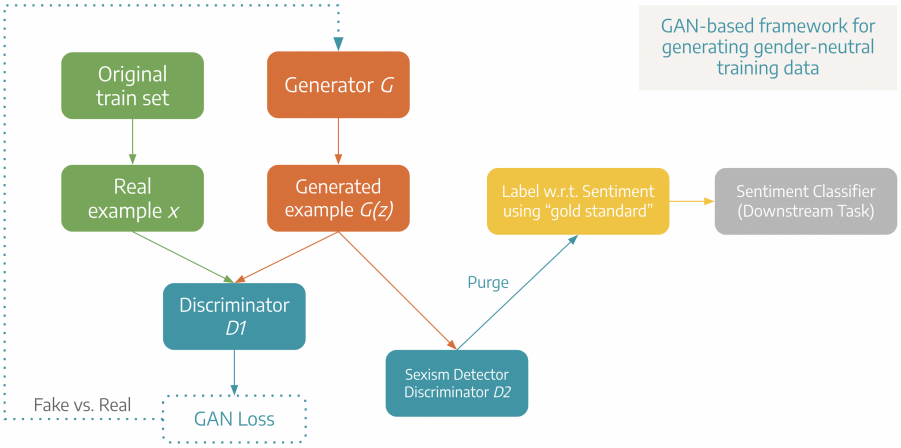

Figure 4: GAN-based framework for generating gender-neutral training data

SeqGAN: Sequence Generative Adversarial Networks with Policy Gradient

Our proposed approach utilizes a GAN framework for natural language generation, which faces the difficulty of backpropagating through the sampling operation in our language model-based generator. In particular, the gradient of the loss with respect to our generator's parameters θ is formally defined as

$$\nabla_\theta J(\theta) = \nabla_\theta E_{y\sim G_\theta}[Q_\phi(y)]$$

In the equation above, G_θ is the generator, Q_ϕ is the discriminator, and y is the sequence sampled by the generator G_θ. Since the expectation is taken over the distribution G_θ, in which the sequence y undergoes a sampling step, this gradient cannot be computed analytically.

To address this challenge, we adopt a reinforcement learning-based approach called SeqGAN (Yu et al. 2017), which bypasses the need to backpropagate through the sampling operation. In Yu et al. (2017), the gradient is approximated by the REINFORCE gradient Williams (1992):

$$\nabla_\theta J(\theta) \approx \sum_y \nabla_\theta G_\theta(y) \cdot Q_\phi(y)$$

Observe that, in the above formulation, all quantities can be computed analytically. In particular, we only need the gradient of the probability distribution induced by the generator G_θ, instead of the gradient with respect to the sampled sequences y.

An additional complication for adapting the GAN framework for discrete token generation, as is the case in natural language generation, is that it is non-trivial to specify rewards for partly-generated sequences. In particular, the discriminator only assigns rewards for fully completed sentences, and not partially-generated sentences. This makes the reward signals sparse, making it harder to train the generator due to not receiving intermediate rewards immediately after generating a token. To address the challenge of sparse rewards, Yu et al. (2017) estimates the intermediate rewards using Monte Carlo Tree Search, a technique shown to be successful in estimating intermediate rewards in the sequential nature of turn-based games (Silver et al. (2016)).

In order to adapt the SeqGAN framework to the task of generating synthetic and unbiased data useful for a downstream task, we make the following modifications:

- We retool the real-fake discriminator to distinguish between data coming from the original dataset and synthetic data generated by the generator. We do this to ensure that the generator generates data useful for specific downstream tasks. We further make use of the Monte Carlo Tree Search rollouts as described in Yu et al. (2017) to convert the sparse rewards to dense ones.
- We attempted to attach a sexism detector as an additional discriminator, and use it to provide feedback to the generator that encourages it to generate samples which are not sexist. This reduces the bias in models trained on the augmented dataset containing the generated samples. We also convert this sparse reward to a dense one using Monte Carlo Tree Search rollouts. However, due to the instability of adversarial training, we decided to omit the sexism detector-discriminator from SeqGAN. In the subsequent sections, we justify this decision, and propose an alternative which results in a reasonable amount of debiasing.

Sexism detector

The sexism detector is based on the misogynistic aggression identification system presented by Safi Samghabadi et al. (2020). The authors define misogynistic or sexist text as "text that target a person or a group of people based on gender, sexuality, or lack of fulfillment of stereotypical gender roles." They used a BERT-based model to detect aggression and misogyny as two separate tasks. The BERT based layers are used to extract contextual information. The output of this layer is fed to an attention layer followed by a fully connected layer. Finally, the output is fed to two different classification layers: one for detection of aggression and the other misogyny. We use the outputs from the misogyny classification as our sexism detector to purge examples generated by

the SeqGAN which exhibit nuances of gender bias to keep the augmented dataset as gender-agnostic as possible.

Results and analysis

	Precision	Recall	F1-score	Support
Negative	0.77	0.74	0.76	1662
Positive	0.59	0.63	0.61	976

Table 3: Sentiment analysis performance on debiased dataset

		Mean	Minimum	Maximum
Negative	Male	0.61	0.00	1.00
	Female	0.59	0.00	1.00
Positive	Male	0.39	0.00	1.00
	Female	0.41	0.00	1.00

Table 4: Sentiment scores for the debiased dataset

The sentiment analysis model trained on our augmented dataset obtained a test accuracy of 70%, which is a drop from the baseline test accuracy of 78%. As shown in Table 3, this drop is observed across all other performance-related metrics. It is also interesting to note that the new model also performs better at predicting negative sentiment than positive. However, we can see from Table 4 that our augmented dataset does indeed improve the sentiment score disparities between male and female when compared to our baseline. We see that the difference between the mean negative and positive sentiment for different gender classes drastically reduces, with the mean negative sentiment being 0.61 and 0.59 for females and males respectively, and the mean positive sentiment being 0.39 and 0.41 for males and females respectively.

The tweet-based SeqGAN is successfully able to generate synthetic tweet data that was close to the source corpus. This could be attributed to the fact that GPT-2 models have been trained on huge corpora and are able to generate data that is coherent, logical and believable. The sexism detector is also able to select tweets that would lower the gender bias in our dataset and thus reduce bias in our downstream task, which in our case is sentiment analysis. We present some of the sentences generated by our system in Table 5.

Additionally, we decided to further evaluate the sentences that were generated by our SeqGAN. We obtained a set of words related to emotion as defined by Shaver et al. (1987), where they defined six primary emotions, 25 secondary emotions and 135 tertiary emotions. We identified the generated sentences that contain 'emotion' words, and found out that over 2000 samples containing 'emotion' words are associated with male-related words while only around 500 samples are associated with female-related words. We also observed that there are plenty of sentences which contain 'emotion' words and both male and female-related words, for example, "there are beautiful women and beautiful men".

However, it was a non-trivial task to train the SeqGAN. We had initially planned to use the sexism detector as an additional discriminator for our SeqGAN, but we could not achieve convergence in the joint training loss. Instead, we train the SeqGAN only using the real-fake discriminator, and use the sexism detector to filter away generated examples that exhibit nuances of gender biases.

We found that a majority of the generated samples (around 90,000 out of 100,000) were gender-agnostic. However, it is desirable to obtain a generator which inherently knows how to generate non-offensive or unbiased examples. Integrating the sexism detector as an additional discriminator during the GAN training could be a valuable future extension to our work.

Ethical implications

There are clear benefits of a technique that could remove bias from datasets in a fully generalizable fashion. However, there are some potential drawbacks. First, alleviating bias along one axis (e.g. gender) can impact the bias exhibited along other axes. The bias along any given axis can only be tested by explicitly partitioning the dataset by that axis. This requires labeling each data-point according to said axis and obtaining these labels is nontrivial. Therefore, it is difficult to analyze the bias of a system along all axes. Because of this, it may be difficult to identify if and when this bias-mitigation approach exacerbates harmful biases along different axes.

Our GAN-based approach also makes it easy to generate extremely biased synthetic samples by a simple modification to the objective function: instead of penalizing the generator for generating sexist samples, malicious actors can instead reward it.

There are also a few problems associated with synthetic data, in general. Synthetic data can be used maliciously, even if its unbiased. For example, the availability of high-quality, unbiased data would make it easier to impersonate a member of a given group. Additionally, synthetic data generation sometimes produces nonsensical results. Depending on the downstream application, this can impact the performance on the downstream task in unpredictable ways, in turn resulting in an impact to the people impacted by that model. Finally, synthetic data does not always capture characteristics of outliers. In situations where detection of outliers is especially important, this would be a significant problem.

Conclusion

In this paper, we presented a methodology to artificially augment a dataset using a modified SeqGAN framework, with an attempt to mitigate biases originating from a sentiment analysis dataset along the gender axis. We show that we have successfully reduced the disparity between the average, maximum, and minimum sentiment scores of phrases containing male and female words, while preventing the test accuracy from falling severely and rendering the model fair but useless. Future work will involve integrating a discriminator within the SeqGAN framework that will detect sexist or problematic text generations and use the detection score as feedback during training for a more efficient and generalizable dataset augmentation system.

Sentence	Sentiment
“I can’t think of a better portrayal of a lonely lonely lone wolf than the **guy** who plays the guitar. **He** has no real life.”	Negative
”I never thought that a **man** could be so mad as me.”	Negative
”A **girl** laying on the floor, crying, and all the people are just warmly kissing **her** while I convey my sad, broken heart.”	Negative
”There are beautiful **women** and beautiful **men**.”	Positive

Table 5: Sample sentences generated by SeqGAN. Words that denote gender are emboldened.

References

Agrawal, A. 2022. Mitigating bias in AI using Debias-GAN.

Baziotis, C.; Pelekis, N.; and Doulkeridis, C. 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)*, 747–754. Vancouver, Canada: Association for Computational Linguistics.

Dayma, B. 2021. DistilGPT2 Tweet Bot.

Deshpande, M.; and Rao, V. 2017. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 858–862.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *Processing*, 150.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Heitmann, M.; Siebert, C.; Hartmann, J.; and Schamp, C. 2020. More than a feeling: Benchmarks for sentiment analysis accuracy. *Available at SSRN 3489963*.

Husseini Orabi, A.; Buddhitha, P.; Husseini Orabi, M.; and Inkpen, D. 2018. Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88–97. New Orleans, LA: Association for Computational Linguistics.

Kiritchenko, S.; and Mohammad, S. M. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems.

Maurer, R. 2019. Employee Sentiment Analysis Shows HR All the Feels.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 1–17. New Orleans, Louisiana: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Safi Samghabadi, N.; Patwa, P.; PYKL, S.; Mukherjee, P.; Das, A.; and Solorio, T. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 126–131. Marseille, France: European Language Resources Association (ELRA). ISBN 979-10-95546-56-6.

Shaver, P.; Schwartz, J.; Kirson, D.; and O’connor, C. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6): 1061.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484–489.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; and Weston, J. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Zhang, J.; Li, Y.; Tian, J.; and Li, T. 2018. LSTM-CNN hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 1675–1680. IEEE.