Example Templates
You are a *(adjective) (identity term).*
(verb) (identity term).
Being *(identity term)* is *(adjective)*
I am *(identity term)*
I hate *(identity term)*

Table 2: Example of templates used to generated an unbiased test set.

Type	Example Words
Offensive	disgusting, filthy, nasty, rude, horrible, terrible, awful, worst, idiotic, stupid, dumb, ugly, etc.
Non-offensive	help, love, respect, believe, congrats, hi, nice, great, fun, nice, neat, happy, good, best, etc.

Table 3: Example of offensive and non-offensive verbs & adjectives used for generating the unbiased test set.

lary (See Table 3) to retain balance in neutral and abusive samples.

For the evaluation metric, we use 1) AUC scores on the original test set (Orig. AUC), 2) AUC scores on the unbiased generated test set (Gen. AUC), and 3) the false positive/negative equality differences proposed in Dixon et al. (2017) which aggregates the difference between the overall false positive/negative rate and gender-specific false positive/negative rate. False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) are defined as below, where $T = \{male, female\}$.

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

Since the classifiers output probabilities, equal error rate thresholds are used for prediction decision.

While the two AUC scores show the performances of the models in terms of accuracy, the equality difference scores show them in terms of fairness, which we believe is another dimension for evaluating the model's generalization ability.

4.2 Experimental Setup

We first measure gender biases in `st` and `abt` datasets. We explore three neural models used in previous works on abusive language classification: Convolutional Neural Network (CNN) (Park and Fung, 2017), Gated Recurrent Unit (GRU) (Cho et al., 2014), and Bidirectional GRU with self-attention (α-GRU) (Pavlopoulos et al., 2017), but with a simpler mechanism used in Felbo et al. (2017). Hyperparameters are found using the validation set by finding the best performing ones in terms of original AUC scores. These are the used hyperparameters:

1. CNN: Convolution layers with 3 filters with the size of [3,4,5], feature map size=100, Embedding Size=300, Max-pooling, Dropout=0.5

2. GRU: hidden dimension=512, Maximum Sequence Length=100, Embedding Size=300, Dropout=0.3

3. α-GRU: hidden dimension=256 (bidirectional, so 512 in total), Maximum Sequence Length=100, Attention Size=512, Embedding Size=300, Dropout=0.3

We also compare different pre-trained embeddings, `word2vec` (Mikolov et al., 2013) trained on Google News corpus, `FastText` (Bojanowski et al., 2017)) trained on Wikipedia corpus, and randomly initialized embeddings (*random*) to analyze their effects on the biases. Experiments were run 10 times and averaged.

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

4.3 Results & Discussions

Tables 4 and 5 show the bias measurement experiment results for `st` and `abt`, respectively. As expected, pre-trained embeddings improved task performance. The score on the unbiased generated test set (Gen. ROC) also improved since word embeddings can provide prior knowledge of words.

However, the equality difference scores tended to be larger when pre-trained embeddings were