Transfer Learning in Abusive Language Classification

Jay Shin, Ji Ho Park

Background and Related Work

Overview

What is Abusive Language?

- Definition: Any type of language that could hurt others
- Categories
  - Profanity
  - Hate Speech (sexism, racism, homophobism)
  - Threat
  - Obscene (sexual harassment)
  - Insult
  - Negative Behaviors (rude & disrespectful sarcasm or criticism)
  - Aggression
  - Cyberbullying

Overview

Why is it meaningful to classify Abusive Language?

- Automatic Online Content Moderation
- Prevent Cyber-bullying and Hate Speech

Abusive Language Datasets

Dataset Descriptions:

Sexist/Racist Tweets dataset (Waseem et al., 2016)

- Size: 18K Tweets
  - None: 12k
  - Racist: 2k
  - Sexist: 4k
- Labels: Multi-label (None, Sexism, Racism) => Binary (None, Abusive)
- Collection: Keywords based on list of criteria based on critical race theory
  - Twitter API
  - Followed some prolific users
- Annotation: Experts Annotate
  - Preferred to annotate as sexist for disagreements

Abusive Language Datasets

Dataset Descriptions:

Wikipedia Attack dataset (Wulcyzn et al., 2017)

- Size: 115k Wikipedia Comments from Wikipedia Article Discussion pages
  - 11.7% positives (most came from blocked users)
- Labels: Binary (Attack or not)
- Collection: Random comments + Sampling comments from blocked users
- Annotation: Crowdsourcing from Crowdflower platform

Abusive Language Datasets

Dataset Descriptions:

Abusive Tweets dataset (Founta et al., 2018)

- Size: 60k
  - Around 20% positive
- Labels: Abusive/Hateful vs None/Spam
- Collection: Twitter Stream API
- Annotation: Crowdsourcing, but very iterative and systematic making the process reliable among others

Previous Approaches in Abusive Language

Mainly focused on performance measured on single datasets.

- Waseem et al. SVM / LR with N-grams on **Sexism/Racism Tweets** (dataset paper)
- Park et al. Hybrid CNN / Word CNN / Char CNN on **Sexism/Racism Tweets**
- Wulczyn et al. Word/Char based LR / MLP on **Wikipedia Attack** (dataset paper)
- Pavlopoulos et al. Deep Attention with GRU on **Wikipedia Attack**

Insights from previous work

- Small Dataset
- Overfitting (generalization is hard)
- Unbalanced Data
- Noisy Data due to questionable data collection and annotation methods.
- Hard to collect and annotate new data from different domain

Problem Statement

- We want to mitigate these problems by transferring knowledge from large dataset to a smaller one: from Wikipedia (115k) to Sexism/Racism Twitter (18k) or Abusive Twitter (60k)

1. Can we use a Wikipedia dataset to train an abusive language classifier for tweets?
2. Can we make a more generalizable model using limited training dataset?

1. Domain Adaptation

Can we use a Wikipedia dataset to train an abusive language classifier for tweets?

What is Domain Adaptation, and Domains?

- Domains here are not what we use in math, but rather data distributions.
- **Domain Adaptation** is a branch of Transfer Learning, in which we aim to learn from a source data distribution and transfer knowledge to a related but different target data distribution.

Domains in Sentiment Analysis

In sentiment analysis of Amazon product reviews, different domains are reviews from different categories of products:

- E.g. *Infantile* is a bad sentiment in other domains like movie reviews or electronics, but it is neutral in baby products.
  - *… an infantile stroller …*
  - *… the movie was surprisingly infantile ...*

Domains in Abusive Language

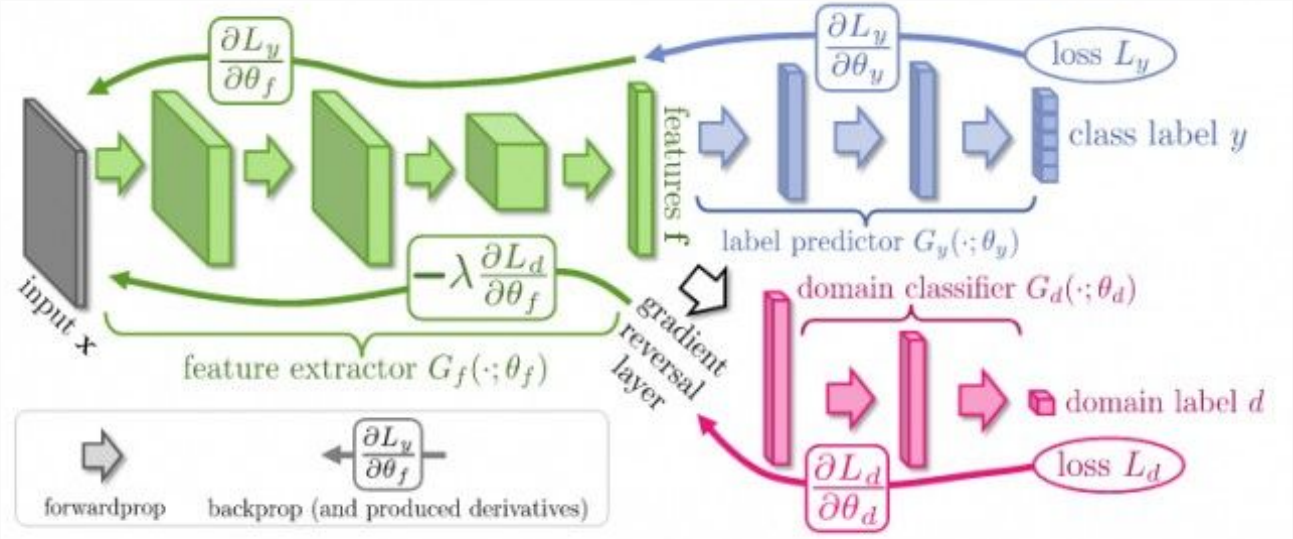
Naturally, we hypothesize that abusive language in different social media (e.g. Wikipedia, Twitter, Facebook, News, etc.) will show different behaviors, but with some generality.

Related Work in Domain Adaptation

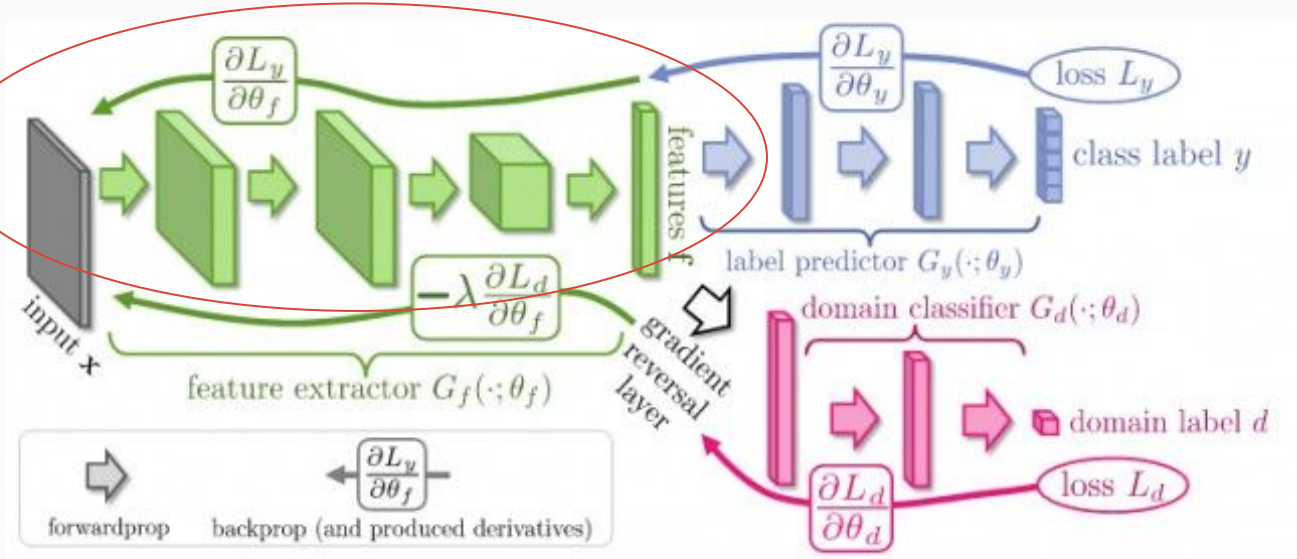
- Ben-David et al. 2007 - Upper Generalization Bounds on domain adaptation
- Ajakan et al. 2014 - Domain Adversarial Neural Networks (DANN) -> Sentiment Analysis on Amazon Reviews & Image Classification
- Bousmalis et al. 2016 - Domain Separation Networks (Image Classification)
- Liu et al. 2017 - Adversarial Shared Private Multi-Task Learning (Sentiment Analysis on Amazon Reviews)

Methodology

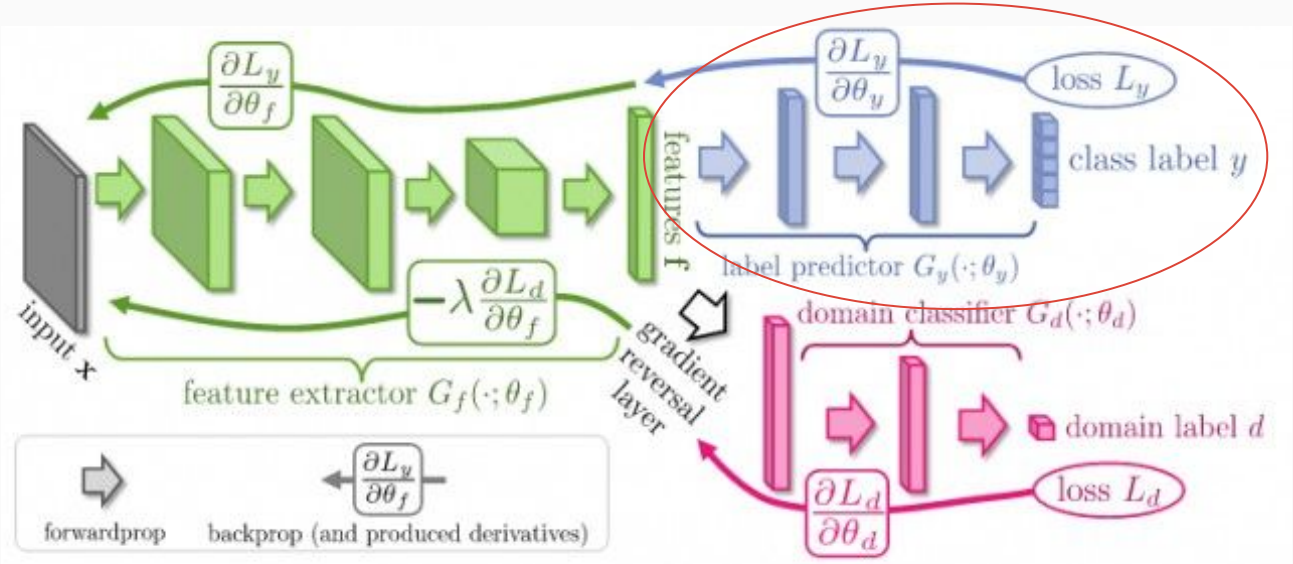
Domain Adversarial Neural Network (Ganin et al, 2015)



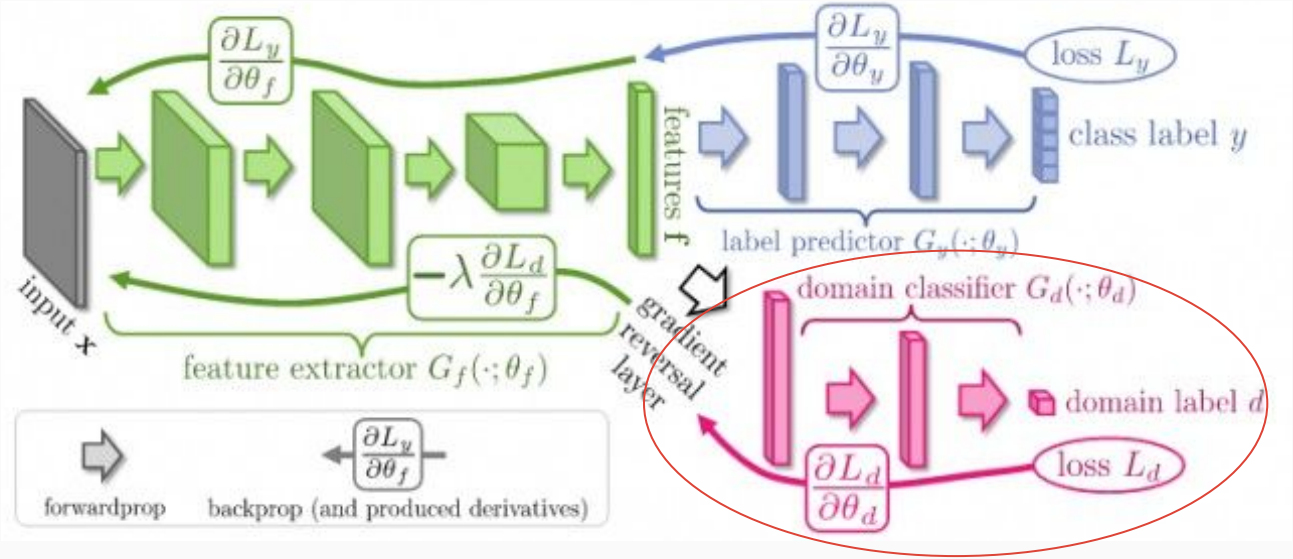
Domain Adversarial Neural Network (Ganin et al, 2015)



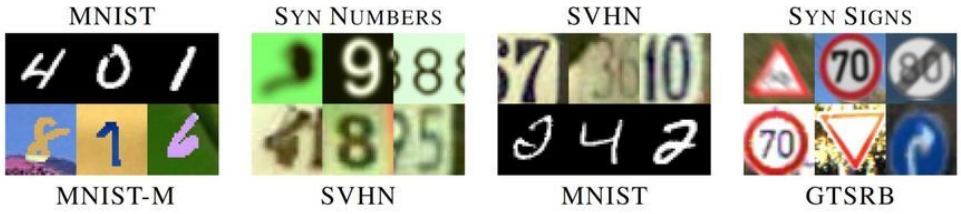
Domain Adversarial Neural Network (Ganin et al, 2015)



Domain Adversarial Neural Network (Ganin et al, 2015)



Domain-adversarial training



METHOD	SOURCE	MNIST	SYN NUMBERS	SVHN	SYN SIGNS
	TARGET	MNIST-M	SVHN	MNIST	GTSRB
SOURCE ONLY		.5749	.8665	.5919	.7400
SA (FERNANDO ET AL., 2013)		.6078 (7.9%)	.8672 (1.3%)	.6157 (5.9%)	.7635 (9.1%)
PROPOSED APPROACH		**.8149** (57.9%)	**.9048** (66.1%)	**.7107** (29.3%)	**.8866** (56.7%)
TRAIN ON TARGET		.9891	.9244	.9951	.9987

Lower Baseline

Upper Baseline

Yaroslav Ganin, Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, ICML, 2015

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Domain-Adversarial Training of Neural Networks, JMLR, 2016

Experiment Settings

Source	Label	**Target**	Label
Wikipedia Attack	None vs. Attack	Abusive Tweets	None/Spam vs. Abusive/Hateful
		Sexist/Racist Tweets	None vs. Sexist/Racist

Goal: training a classifier with wikipedia train set (labeled)/twitter train set (unlabeled) and test on twitter test set (labeled)

Baseline experiment results

Wikipedia and Twitter (F1 score)

Lower baseline	Train on wikipedia train set and evaluate on twitter valid set	0.762
Upper baseline	Train on twitter train set and and evaluation on twitter valid set	0.779

- Such a small margin. It means that the data distribution is very similar except for the length! And RNN and attention model effectively handles the length variation
- **Domain adaptation training did not work well.**
- Our assumption that wikipedia/twitter abusive language are separate domains may be wrong

Top words with Tf-idf scores in positive samples

'fuck','fucking','number',**'shit'**,'bitch',’like','wikipedia',**'ass'**,'suck',**'stupid'**,'stop','just','asshole','page',**'idiot'**,'gay','dick','know','faggot','cunt','life','people','did',**'hell'**,'little','hey','think','block','want','dont','cock','shut','die','talk','time',**'hate'**,'going','ll','piece','edit','really','right','article','delete','make','fag',blocked',

WikiDetox

'user','rt',**'fucking'**,'url',**'hate'**,**'idiot'**,**'ass'**,**'bitch'**,like',**'fuck'**,'bad','number',**'stupid'**,'im','just','dont','idiots','shit','people','damn','ugly',**'hell'**,'know','sick','bitches','trump','nigga','niggas','fucked',’think','got','annoying','say','worst','disgusting',’really','look',‘time','fuckin','stop','man','retarded','going','make','hes','nasty','did',

Founta(twitter)

On the other hand, if we use sexist/racist dataset

Wikipedia and Twitter(waseem: sexist+racist merged) (F1 score)

Lower baseline	Train on wikipedia train set and evaluate on twitter valid set	0.294
Upper baseline	Train on twitter train set and and evaluation on twitter valid set	0.710

- This time, such a big margin!
- **Domain adaptation training did not work well.**
- Maybe tasks of classifying sexist and/or racist language and classifying abusive language on wikipedia is too different to use domain adaptation

Top words with Tf-idf scores in positive samples

'fuck','fucking','number','shit',**'bitch'**,'like','wikipedia','ass','suck','stupid','stop','just','asshole','page','idiot','gay','dick','know','faggot','cunt','life','people','did','hell','little','hey','think','block','want','dont','cock','shut','die','talk','time','hate','going','ll','piece','edit','really','right','article','delete','make','fag',blocked',

WikiDetox

'rt','sexist','women','islam','http','mkr','just','kat','muslims','like','girls','notsexist','men','female','muslim','woman','mohammed','think','know','people','ca','does','want','hate','religion','maxblumenthal','mt8_9','prophet','jews','did','quran','isis','girl','really','football',
'im','right',**'bitch'**,'sports','good','israeliregime','man',

Waseem(twitter)

2. Tackling Bias

Can we make a more generalizable model using limited training dataset?

Unintended biases

Discriminative model ***intentionally*** capture biases and inject those biases in the representations so that the model can classify samples well according to the supervised training data.

a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others. (Dixon et al., 2017)

Ex. *I am a gay man*

Why?

Dixon et al, 2017, Zhang et al, 2018 says:

1. Imbalance in data
2. Biased Resources (Word embedding)

Hurts the generalization ability of the model!!

Measuring Biases in the dataset & classifier

We know there is a bias in the dataset and the classifier trained from it.

But how do we measure it?

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Dixon et al., 2017, Measuring and Mitigating Unintended Bias in Text Classification

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

$$\textit{False Positive Equality Difference} = \sum_{t \in T} |FPR - FPR_t| \quad (1)$$

$$\textit{False Negative Equality Difference} = \sum_{t \in T} |FNR - FNR_t| \quad (2)$$

"You are a disgusting woman" vs. *"You are a disgusting man"*

"Being male is good" vs. *"Being female is good"*

Generated unbiased test set with **1152** samples (each half about each gender)

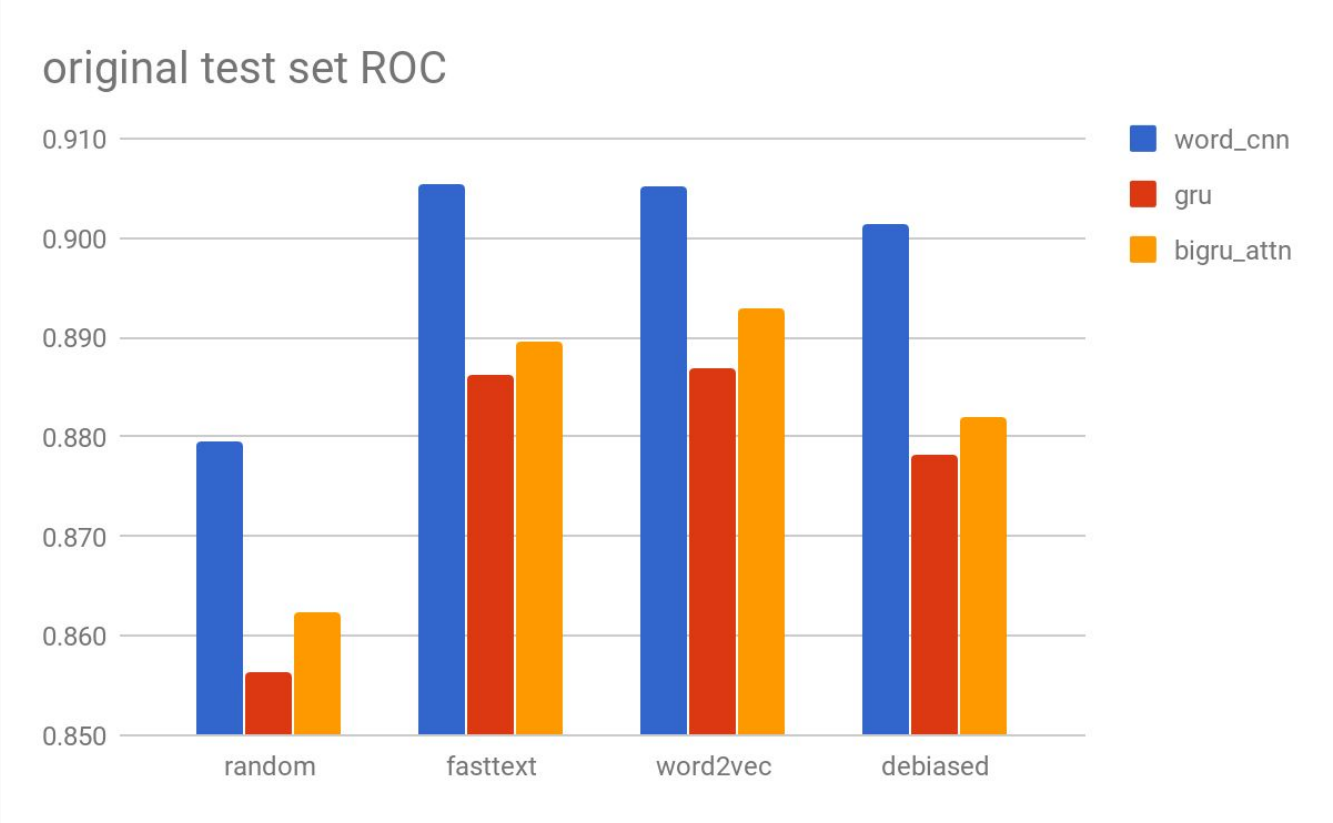
Let's look at False Positive Equality Difference to decide whether the classifier is biased toward gender or not.

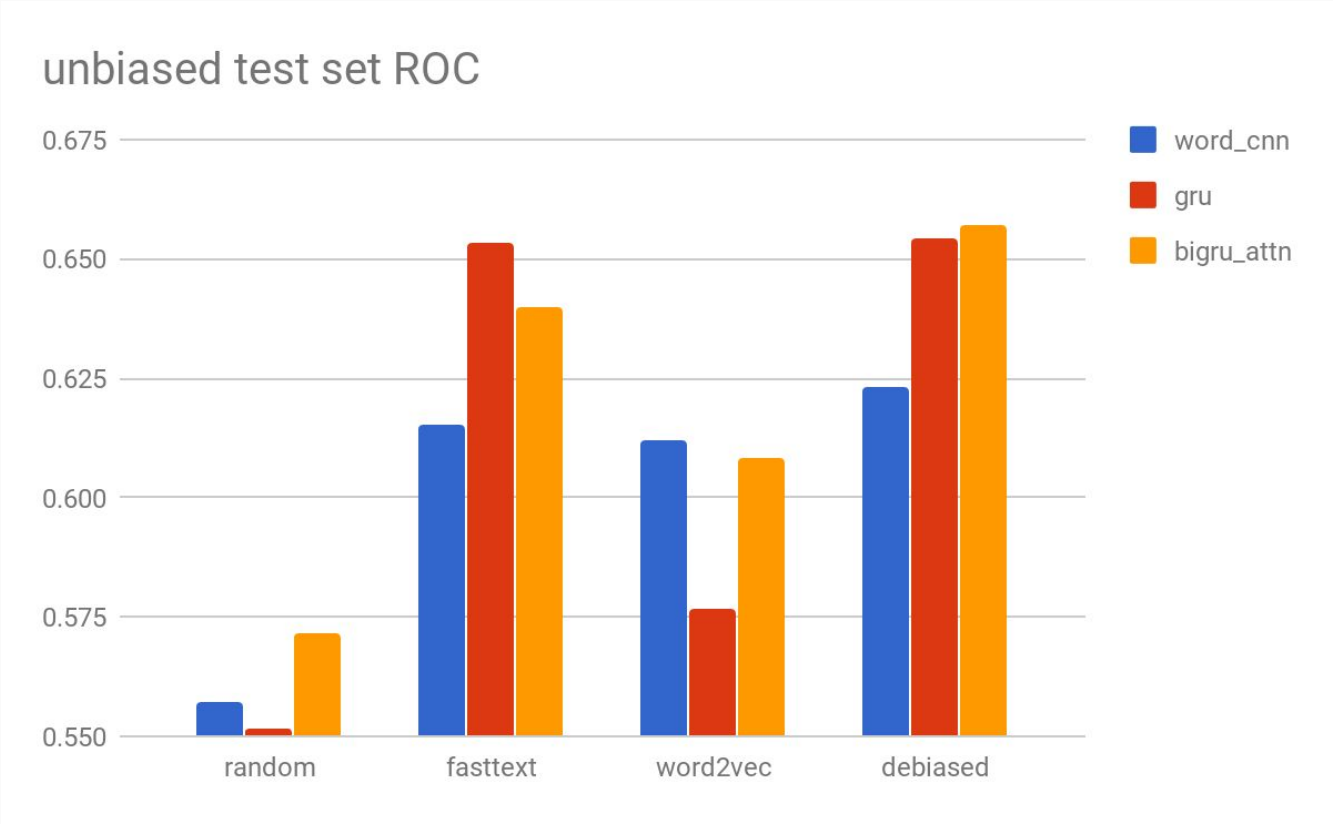
Comparisons

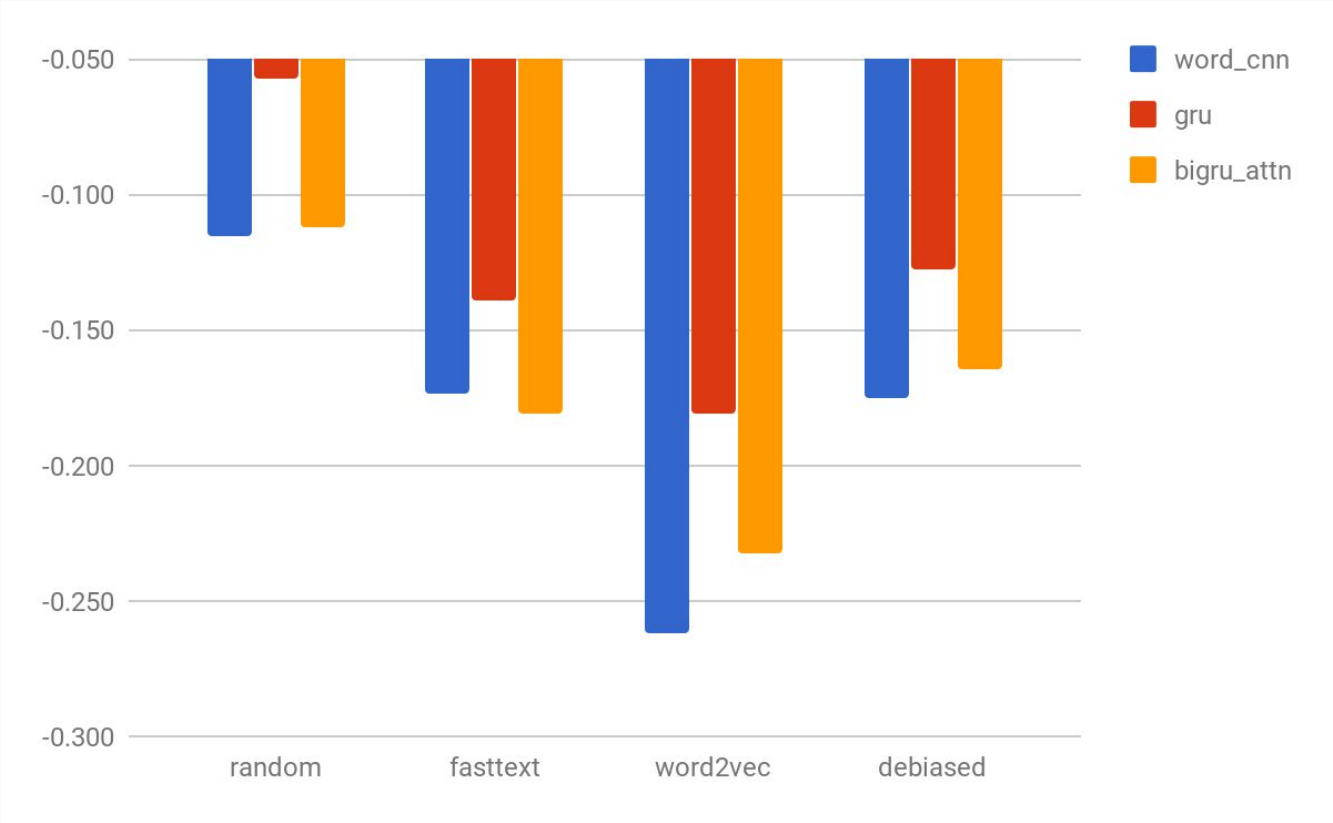
1. Different pretrained embeddings
    a. Bolukbasi et al., 2016 points out gender biases in pretrained embeddings and propose debiased word2vec
    b. Compare random/fasttext/word2vec/word2vec_debiased
2. Different models
    a. Word CNN
    b. GRU
    c. Bidirectional GRU with attention (Pavlopoulos et al., 2017)
3. Different Twitter Datasets
    a. Waseem (none vs. sexist binary classification)
    b. Founta (none/spam vs. abusive/hateful)

All the experiments were run 10 times and averaged

Analysis on Twitter Sexist/~~Racist~~ Dataset (None vs. Sexist/~~Racist~~)


original test set ROC
0.910
0.900
0.890
0.880
0.870
0.860
0.850
random
fasttext
word2vec
debiased
word_cnn
gru
bigru_attn

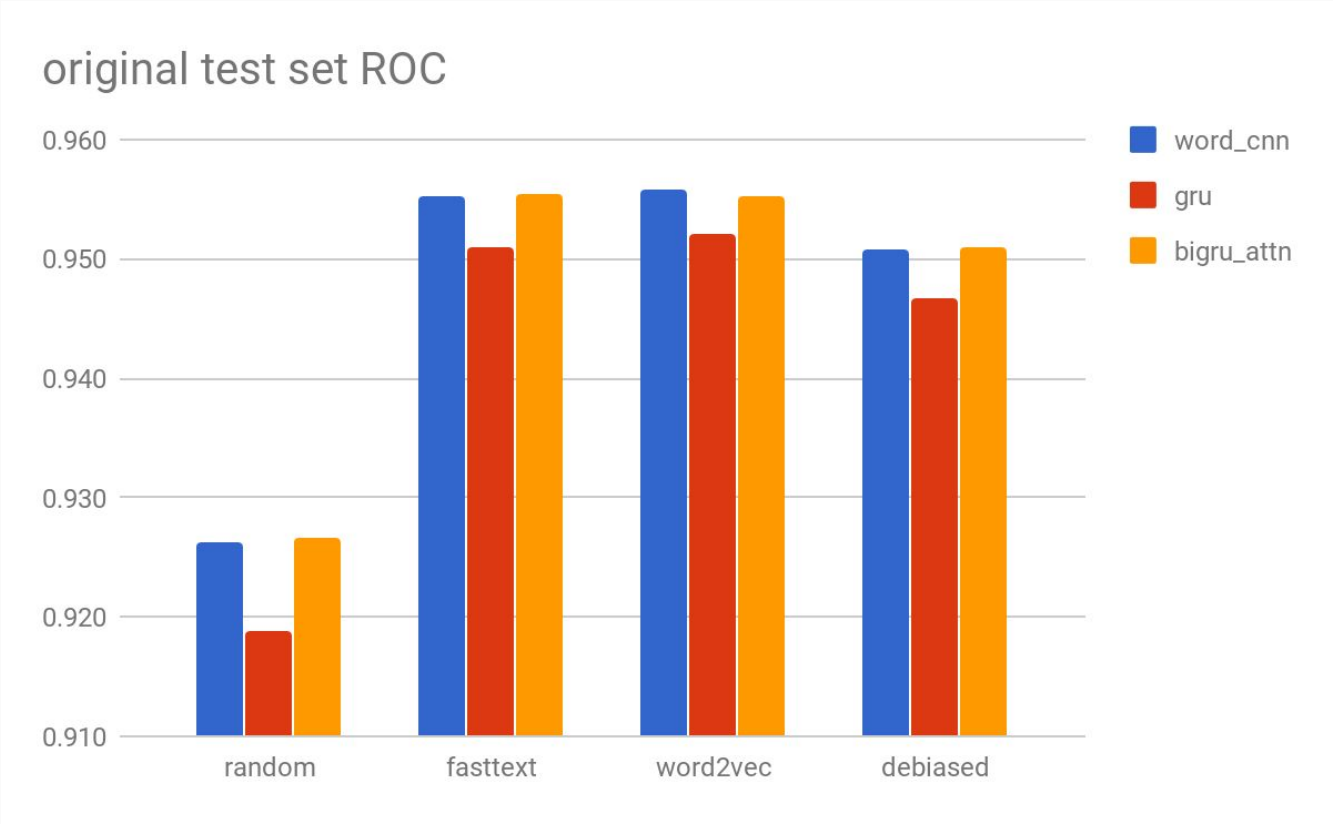

unbiased test set ROC
0.675
0.650
0.625
0.600
0.575
0.550
random
fasttext
word2vec
debiased
word_cnn
gru
bigru_attn



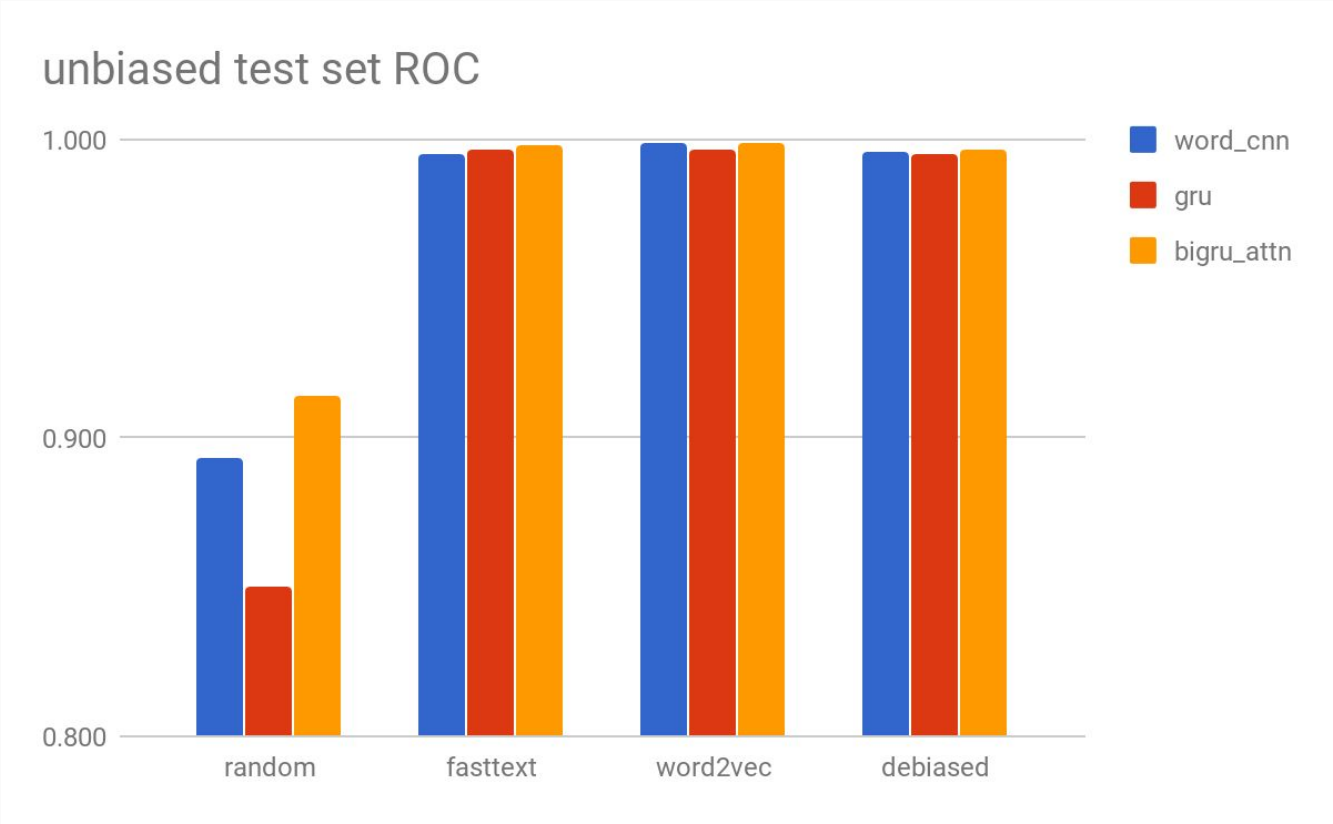
False Positive Equality Difference (woman)

False positive rate (overall) - false positive rate (woman) for unbiased test set

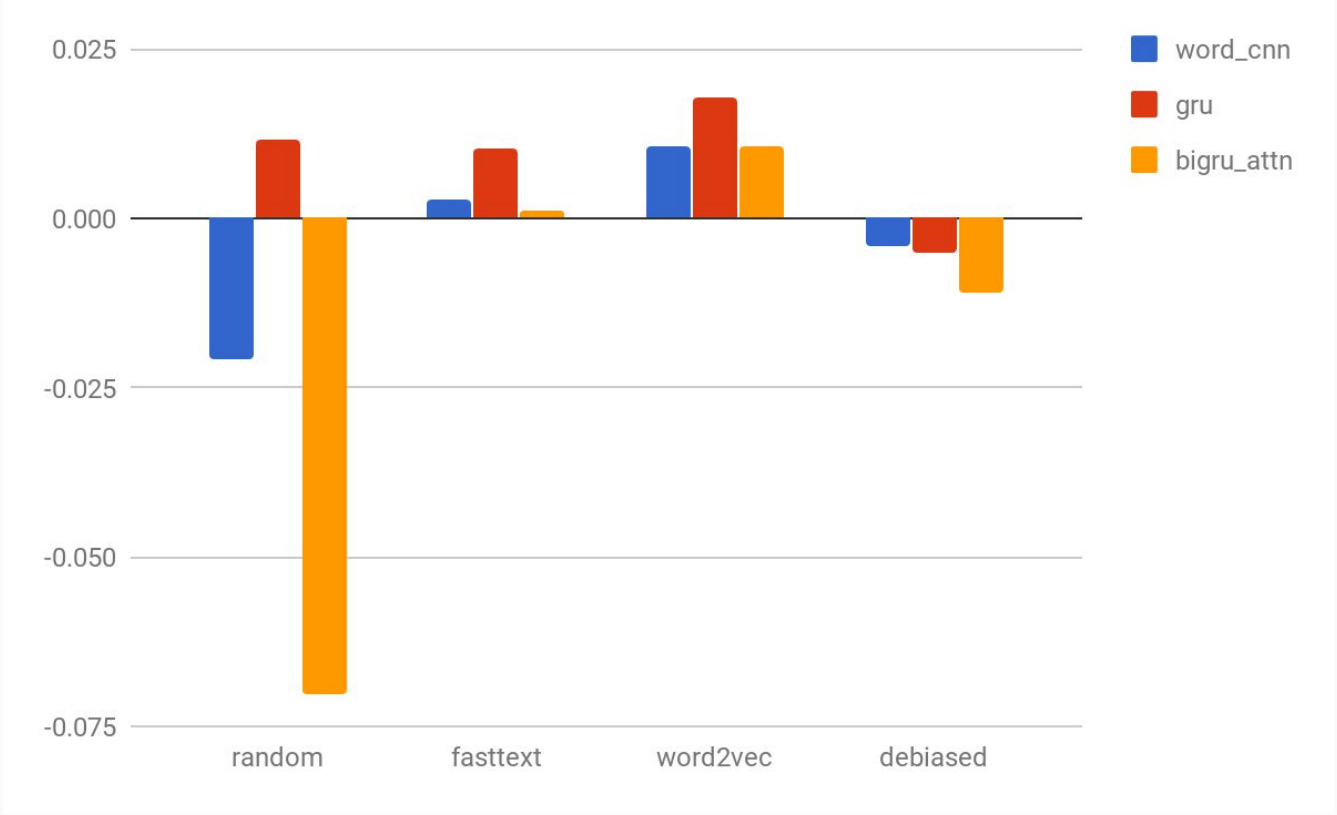
Analysis on
Twitter Abusive Dataset
(None/spam vs. Abusive/Hateful)



Founta (None/spam vs. Abusive/Hateful)



Founta (None/spam vs. Abusive/Hateful)



False Positive Equality Difference (woman)
False positive rate (overall) - false positive rate (woman)
for unbiased test set

Discussions

1. *Dataset size and imbalance* can cause unintended biases toward a certain gender. Due to this, some dataset may be less biased than another dataset. This is important in tasks like abusive language detection.
2. *Pretrained embedding* can push biases into a certain direction, but without them the performance may not be so good. Even debiased embedding does not help depending on the dataset.
3. *Certain model that "attends"* to salient words (like Word CNN's max-pooling /w RNN with attention) can capture not only intended biases (good for performance) but also the unintended bias.

Performance vs. unintended bias trade-off?

How to solve this problem?

1. **Solve the data imbalance problem**
   Dixon et al., 2017 propose adding more negative samples (neutral texts from wikipedia)
2. **Augment the data by flipping the gender**
   Zhao et al., 2018 (Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods), NAACL 2018 short
3. **Transfer learning?**
   If we train the model with a less biased, larger, but similar dataset together, would the bias in the first dataset reduce? Would the model perform better?
   By sharing a representation together, transfer learning might have a regularization effect against overfitting to the imbalance data! This will make the model to learn a better representation so that the model can generalize well (**hypothetical**)

Conclusion

1. *Definition of Domain* is important when using domain adaptation method. Source and Target data should contain not only enough *common features* but also enough *distinct features*.
2. Classifiers trained with publicly available abusive language datasets can contain *unintended biases*. Measuring these biases and mitigating them should also be considered important to build a model with better *generalization ability*.