Null It Out

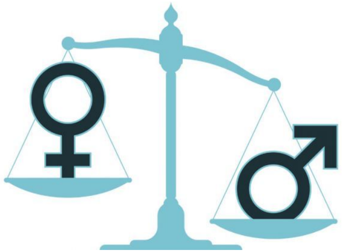
Team 5: Boney M.

Assem, Bauyrzhan, Bella, Ern Chern

Problem

Bias Mitigation in Text Classification

Can machine be biased? Machine learning models learn patterns in the biased data.



Gender Bias

e.g. Man is to woman as computer programmer is to homemaker (Sun et al., 2019)

Racial Bias

e.g. Black is to criminal as white is to police (Manzini et al., 2019)

Age Bias

e.g. Keywords related to older age more likely to be classified as negative (Diaz et al., 2018)

Why is it important to mitigate bias in ML-based classification?

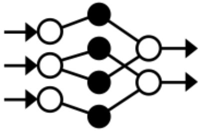
Biased models can enter real-world settings and magnify existing inequality.

Related Work

Bias Mitigation in Text Classification

Debiasing datasets

- Modify biased training datasets
- Problem:
  - Costly manual annotation
  - Need to retrain
- Reweighting datapoints, e.g. (Wang et al., 2019)



Debiasing models

- Modify the word representations
- Zero out components in presupposed bias feature space, e.g. (Bolukbasi et al., 2016)
  - Problem: Non-generalizable
- Apply adversarial training, e.g. (Xie et al., 2017)
  - Problem: Notoriously hard to train

Chosen Paper (ACL 2020)

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] Yanai Elazar[1,2] Hila Gonen[1] Michael Twiton[3] Yoav Goldberg[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher

Why choose this paper?

- Generalizable approach
- No retraining

Replication Approach

Iterative Nullspace Projection

Approach: Remove bias features by projecting them onto the Null Space

Suppose **W(X) ➔ Z**,

with **X**: set of features, **Z** : gender/race/age, **W** : classifier

Goal:

Find **P** such that **W(P(X)) = 0**

i.e. classifier **W** can't predict **Z** based on the **P(X)**

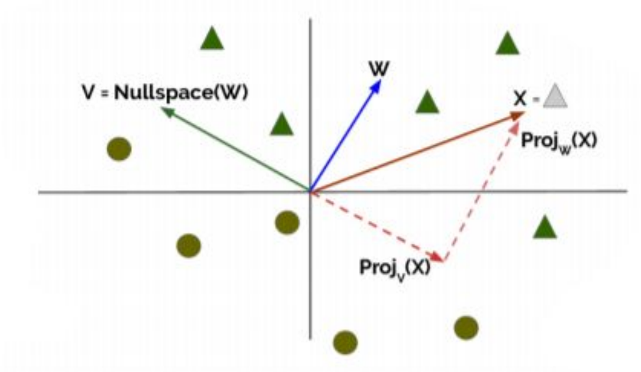


Figure 2: Nullspace projection for a 2-dimensional binary classifier. The decision boundary of W is W 's null-space.

Process:

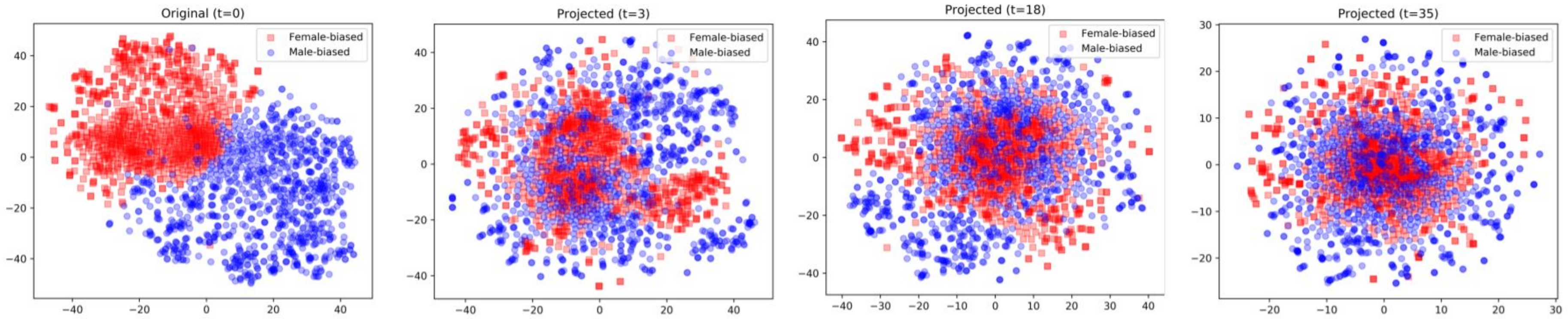
1. Find null space of W
2. Project X onto the null space of W.
3. Now we have protected P(X), where P is the projection matrix.

Replication Approach

Iterative Algorithm

1. Train classifier $\mathbf{W_1}$ on $\mathbf{X}$ and obtain $\mathbf{X_1} = \mathbf{P_1}(\mathbf{X})$
2. Train classifier $\mathbf{W_2}$ on $\mathbf{X_1}$ and obtain $\mathbf{X_2} = \mathbf{P_2}(\mathbf{X_1})$
3. Repeat until no classifier can be trained.

Thus, we removed linear relationships between **Z** and the final projection of **X**.



Application to Fair Classification

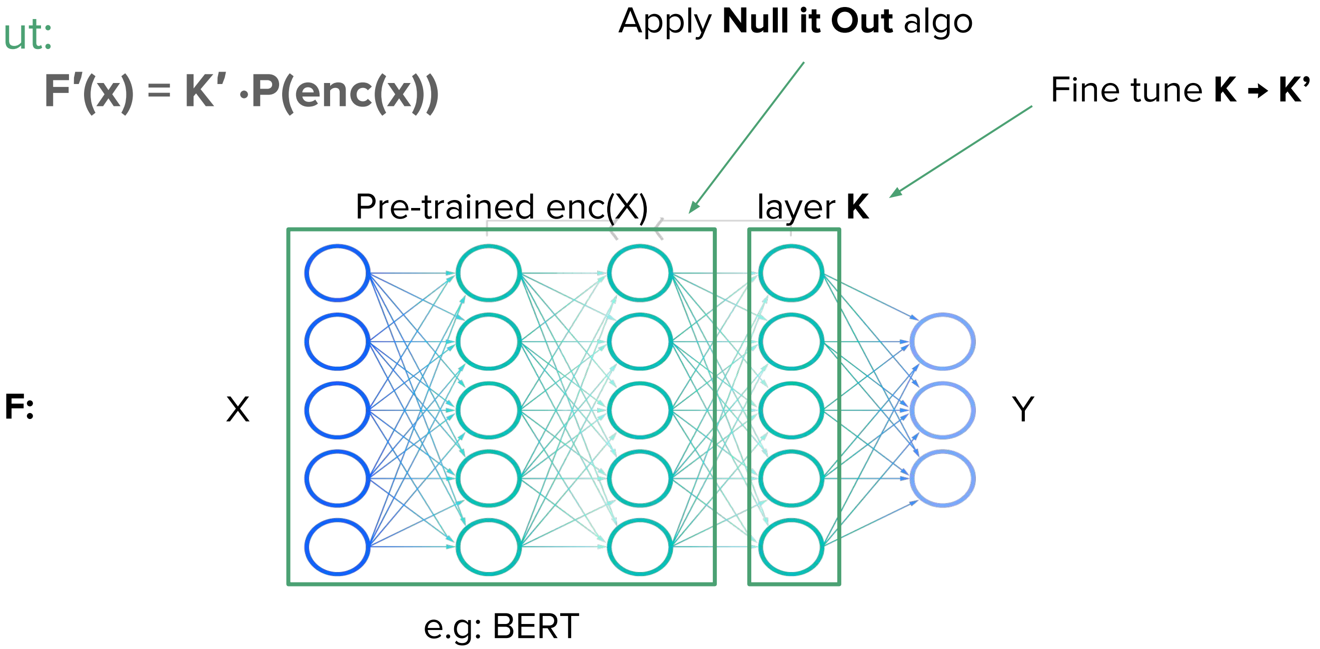
Given:

$$F = K \cdot enc(x)$$

$$F: X \to Y$$

Output:

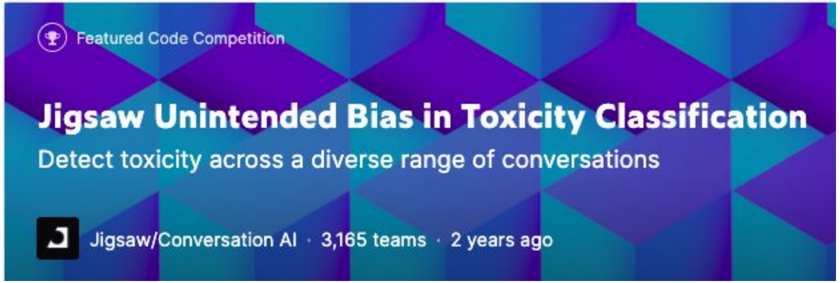
$$F'(x) = K' \cdot P(enc(x))$$



Improvement Approach #1

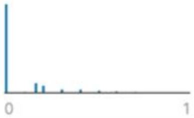
Apply Null it Out on different domains

- Bias in Toxicity detection
    - Gender
    - Ethnicity
    - Race
    - Religion
- Base Model: GRU/LSTM + BERT
- Expected result:
    - Identify unique features of each bias class



Improvement Approach #1

Dataset

comment_text	split	toxicity	male	female	homosexual_gay_or_l...	black	white
1971916 unique values	train 90% test 10%		[null] 78% 0.0 18% Other (88283) 4%	[null] 78% 0.0 18% Other (80984) 4%	[null] 78% 0.0 22% Other (16789) 1%	[null] 78% 0.0 21% Other (21554) 1%	[null] 78% 0.0 21% Other (32908) 2%
"while arresting a man for resisting arrest". If you cop-suckers can't see a problem with this, the...	test	0.8157894736842106					
Tucker and Paul are both total bad ass mofo's.	train	0.55					

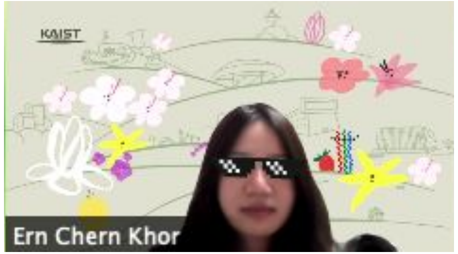
Identity Subgroup	Comment text	Toxic	Toxicity score
Black	Republicans assume all people, including blacks, are capable of having proper ID to vote. Democrats believe blacks are incapable of having proper ID to vote. Who's the racist?	False	79%

Improvement Approach #2

Style Transfer Experiment

Goal: Generate *formal text*

- Assumption: formal text lack emotional features.
- Method: Null out emotional features of the embeddings
- Build text generation model using unbiased embeddings



Improvement Approach #2

Style Transfer Experiment

Goal: Generate *formal text*

Data: SemEval 2018

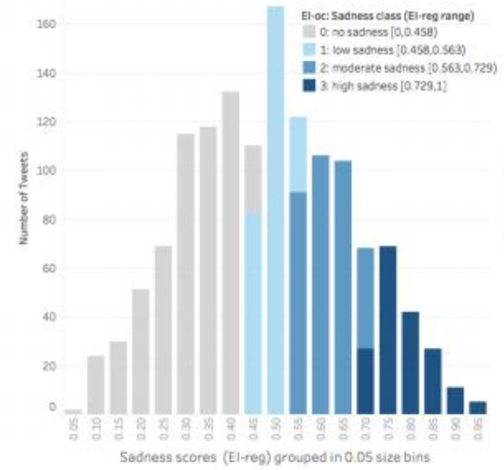
SemEval-2018 Task 1: Affect in Tweets

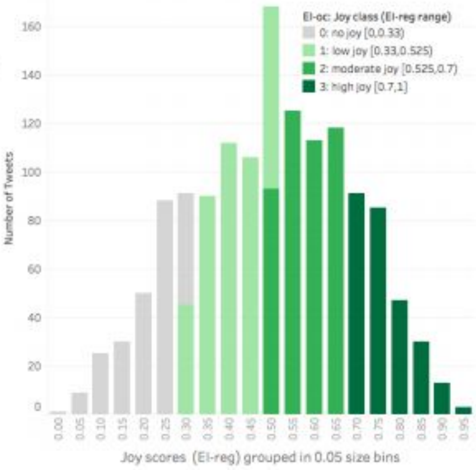
Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Felipe Bravo-Marquez
The University of Waikato, New Zealand
fbravoma@waikato.ac.nz

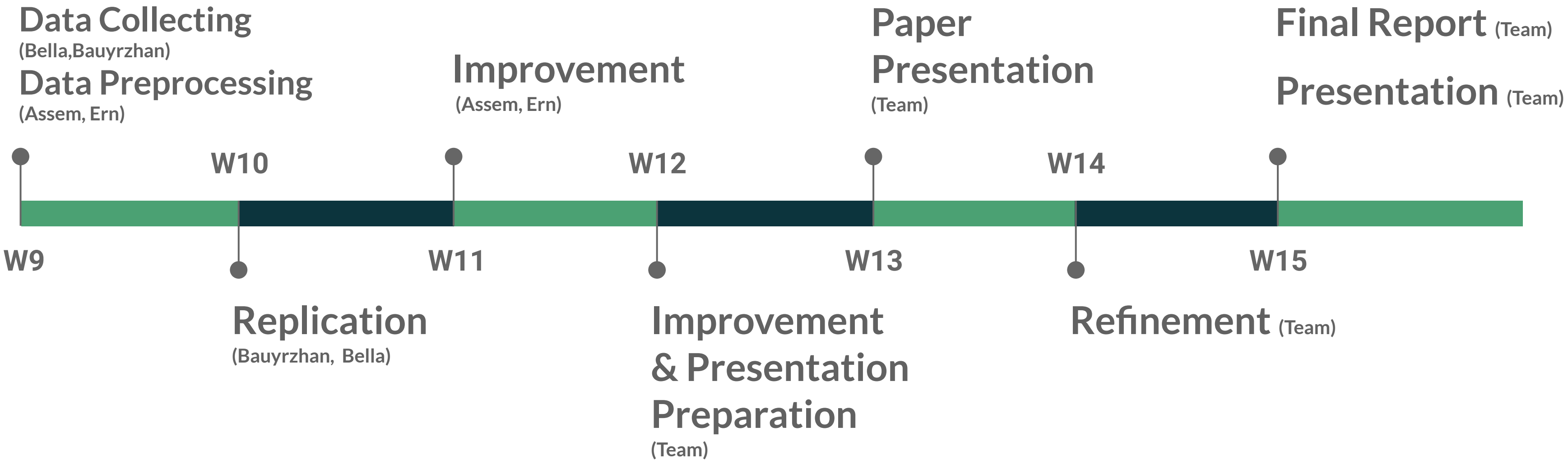
Mohammad Salameh
Carnegie Mellon University in Qatar
msalameh@qatar.cmu.edu

Svetlana Kiritchenko
National Research Council Canada
svetlana.kiritchenko@nrc-cnrc.gc.ca





Weekly Plan



Data Collecting
(Bella,Bauyrzhan)
Data Preprocessing
(Assem, Ern)
W9
W10
Replication
(Bauyrzhan, Bella)
Improvement
(Assem, Ern)
W11
W12
Improvement
& Presentation
Preparation
(Team)
Paper
Presentation
(Team)
W13
W14
Refinement (Team)
Final Report (Team)
Presentation (Team)
W15

Summary

Team 5: Boney M.

- Problem: Bias mitigation in ML models
- Approach: Iterative Null Space Projection
- Improvement:
    - Toxicity detection
    - Style Transfer