Attention-Weighted Integrated Gradients for Target-Aware Cyberbullying Detection

CS329T - Spring 2022

Sharan Ramjee, Sarthak Consul

Introduction

- **Motivation:**
  - Sentiment Analysis for cyberbullying detection systems suffer from robustness issues

- **Dataset:**
  - hatespeech-twitter dataset: ~4.3k examples (63% Normal and 37% Cyberbullying)

- **Our Approach:**
  - twitter-roBERTa-base-sentiment-latest model
  - Attention-Weighted Integrated Gradients
    - Leverage Integrated Gradients completeness property
    - Re-weight Integrated Gradients attributions using aspect-target token self-attention

Attention Weighted Integrated Gradients

- Neutralize usernames

- Attention Weights

$$\alpha_i \begin{cases} = 0 & \text{if } token_i = token_t \\ \propto \max_l \max_h \mathrm{A}(l, h, token_i, token_t) & \text{otherwise} \end{cases}$$

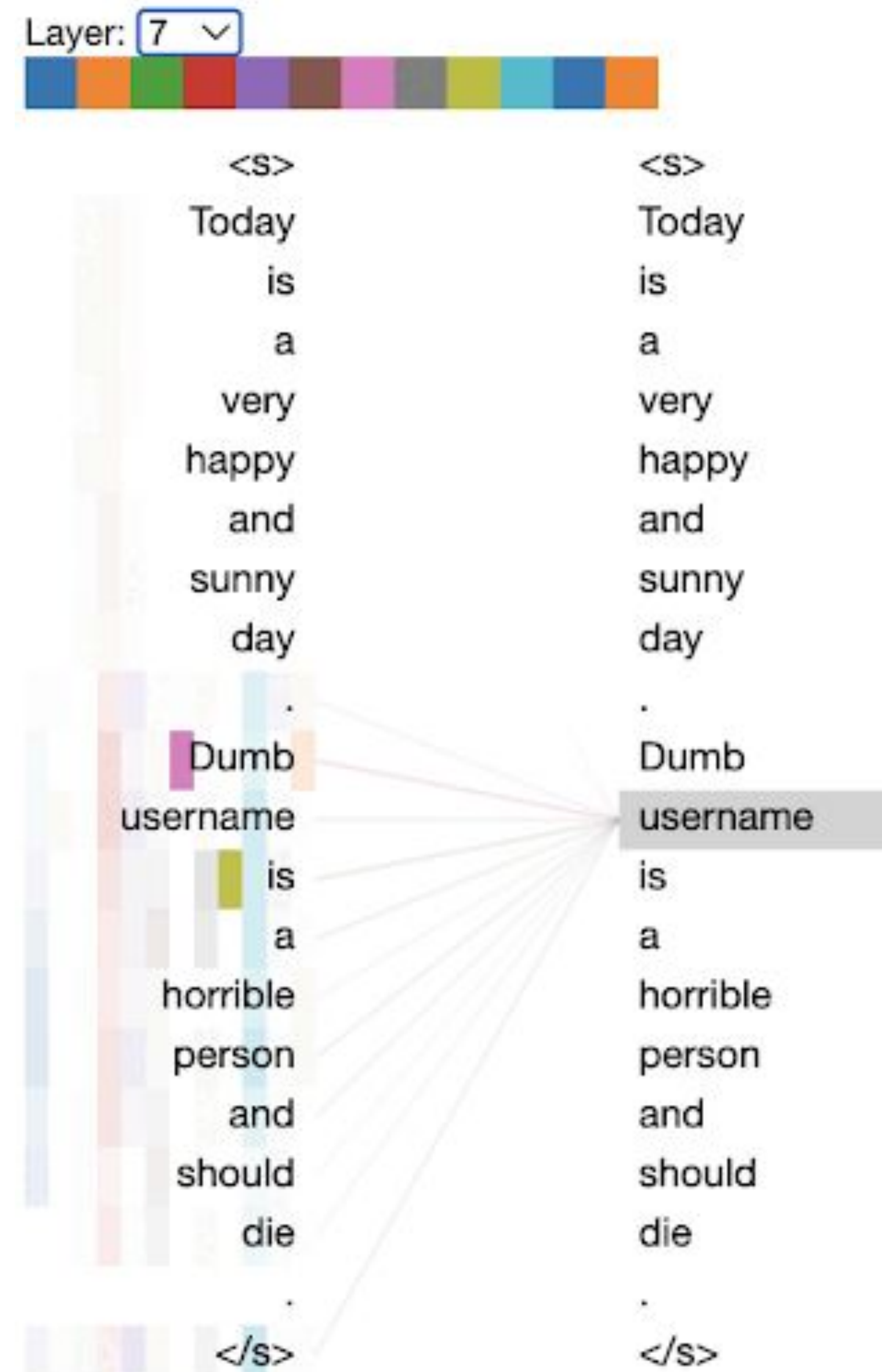
- Attributions via Integrated Gradients

$$QoI = \max(logit_{+ve}, logit_{neutral}) - logit_{-ve}$$

- AWIG Score

$$\text{AWIG Score} = \sum_i \alpha_i \times \mathrm{QoI}(token_i)$$

- Cyberbullying if Score <=0



Stanford University

Robustness Analysis

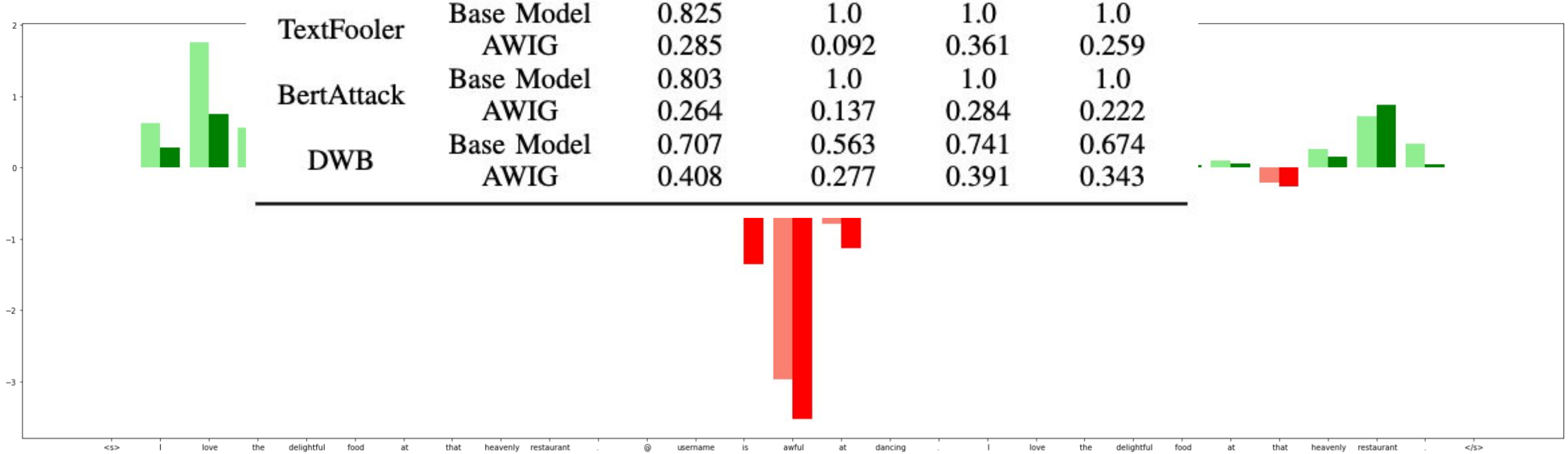
- **Camouflage Attacks**
- **Perturbation Attacks**
  - **Prepend**: Prepend sentences of opposing sentiment to sentence with aspect-target
  - **TextFooler**: Replace important words with synonyms
  - **BertAttack**: Replace important words with those suggested by pre-trained BERT model
  - **DeepWordBu**

TABLE I: Camouflage Attack Analysis

Method	Accuracy	Precision	Recall	F1-Score

TABLE II: Performance Drops on Perturbation Attacks

Attack	Model	Accuracy	Precision	Recall	F1-Score
TextFooler	Base Model	0.825	1.0	1.0	1.0
	AWIG	0.285	0.092	0.361	0.259
BertAttack	Base Model	0.803	1.0	1.0	1.0
	AWIG	0.264	0.137	0.284	0.222
DWB	Base Model	0.707	0.563	0.741	0.674
	AWIG	0.408	0.277	0.391	0.343



Fairness Analysis

- Twitter Usernames
- Protected Attributes
  - Do not want Twitter usernames (can represent protected attributes) to impact model outputs
  - **Race**: African-American English and White-Aligned English
  - … usernames an attribution of 0
  - **Sex**: Female and M…
  - **Political leaning**: F…

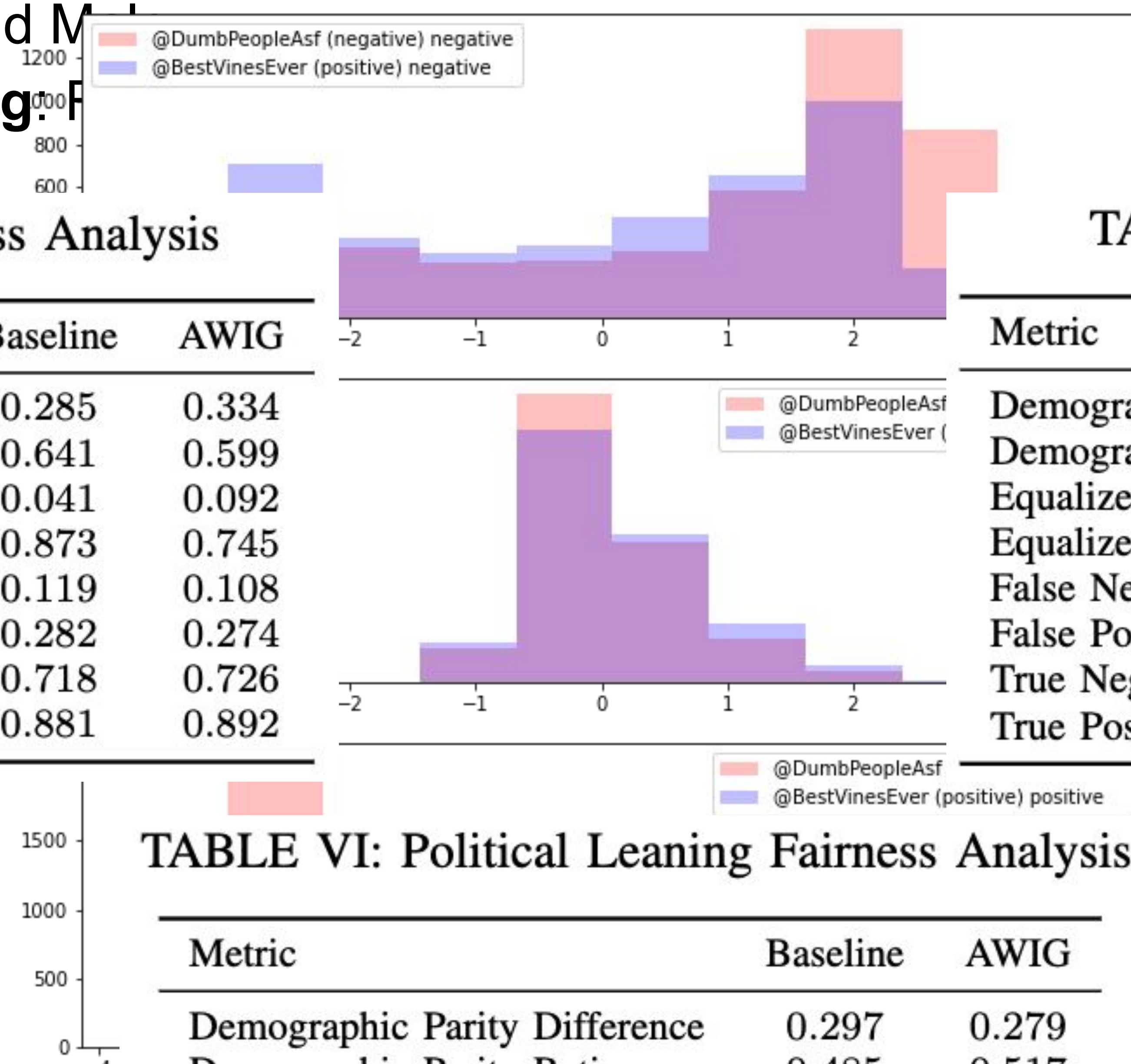


TABLE IV: Race Fairness Analysis

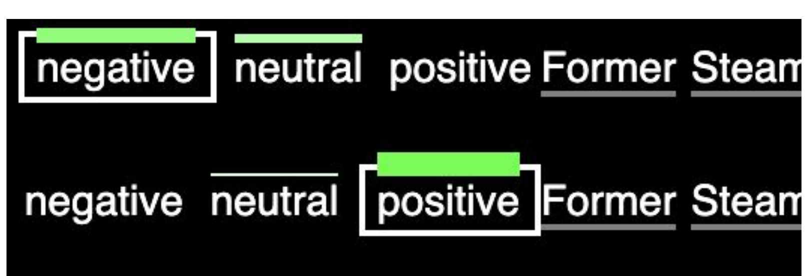
Metric	Baseline	AWIG
Demographic Parity Difference	0.285	0.334
Demographic Parity Ratio	0.641	0.599
Equalized Odds Difference	0.041	0.092
Equalized Odds Ratio	0.873	0.745
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892

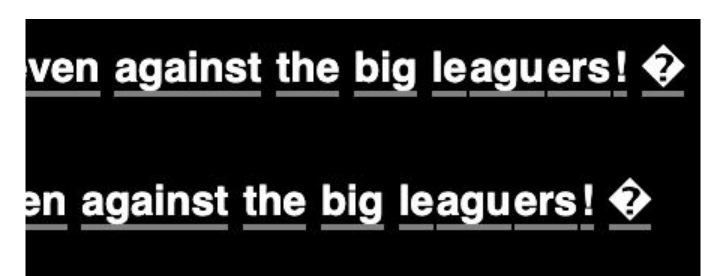
TABLE V: Sex Fairness Analysis

Metric	Baseline	AWIG
Demographic Parity Difference	0.012	0.012
Demographic Parity Ratio	0.979	0.980
Equalized Odds Difference	0.119	0.109
Equalized Odds Ratio	0.881	0.891
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892

TABLE VI: Political Leaning Fairness Analysis

Metric	Baseline	AWIG
Demographic Parity Difference	0.297	0.279
Demographic Parity Ratio	0.485	0.517
Equalized Odds Difference	0.156	0.115
Equalized Odds Ratio	0.464	0.589
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892





Q&A

Code: https://github.com/sharanramjee/cyberbullying-awig