Mitigating Bias in Cyberbullying Detection

CSE 424

Group/Row no. 04

Sec.02

01/ Maliha Binte Masud (21201434)

02/ Abdur Rahman Shafi (23241118)

03/ MD Rakibul Hasan Talukder (23241100)

04/ Amina Zannat Nurhan (23241099)

05/ Sabbir Hossain Mirza (23241086)

06/ Fahmid Hasan Chowdhury (21201286)

07/ Radito Dhali (24141150)

08/ Zakaria Ibne Rafiq (21201357)

Introduction

The social media boom has brought a dark side: cyberbullying. Existing solutions struggle with bias. Our project aims to bridge this gap. We target unbiased detection of online harassment (gender, language, religion, etc.) while achieving high accuracy in classifying harassing text. This aligns with your focus on a diverse dataset for cyberbullying detection.

Data Composition:

- 6 categories: Different cyberbullying types based on age, ethnicity, gender, religion, non-bullying text, and others.
- Balanced distribution: 24,000 entries with even representation across data types.

Dataset Link: Jason Wang, Kaiqun Fu, Chang-Tien Lu, November 12, 2020, "Fine-Grained Balanced Cyberbullying Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/kn1c-zx22.

Methodology

Pre Processing

- Removed null values
- Remove unwanted characters.
- Converting into lower case
- Formatted the raw data (Bag-of-Words)

ML Models:

- Naive Bayes
- Logistic Regression
- K-Nearest Neighbour

Logistic Regression

	precision	recall	f1-score	support
age	0.96	0.97	0.97	812
ethnicity	0.99	0.97	0.98	783
gender	0.89	0.81	0.85	796
not_cyberbullying	0.53	0.50	0.52	819
other	0.54	0.65	0.59	788
religion	0.98	0.94	0.96	802
accuracy			0.81	4800
macro avg	0.81	0.81	0.81	4800
weighted avg	0.81	0.81	0.81	4800

Naive Bayes

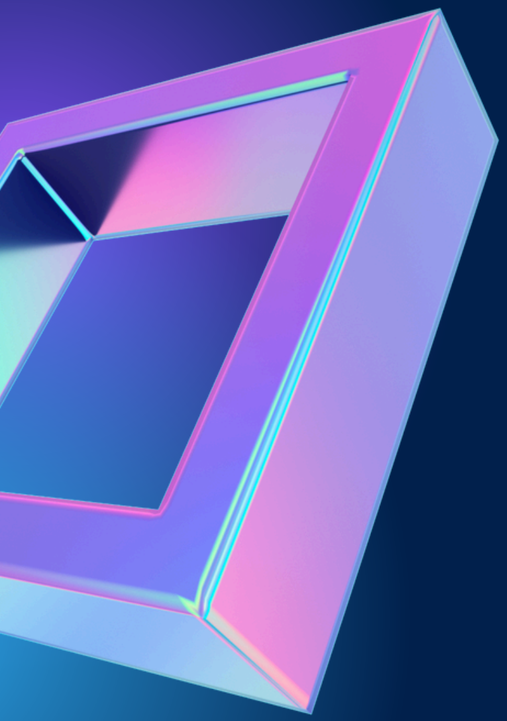
	precision	recall	f1-score	support
age	0.81	0.96	0.88	812
ethnicity	0.78	0.92	0.85	783
gender	0.84	0.77	0.80	796
not_cyberbullying	0.59	0.39	0.47	819
other	0.55	0.51	0.53	788
religion	0.85	0.96	0.90	802
accuracy			0.75	4800
macro avg	0.74	0.75	0.74	4800
weighted avg	0.74	0.75	0.74	4800

K Nearest Neighbour

	precision	recall	f1-score	support
age	0.98	0.96	0.97	812
ethnicity	0.99	0.69	0.82	783
gender	0.97	0.54	0.69	796
not_cyberbullying	0.22	0.06	0.10	819
other	0.27	0.89	0.41	788
religion	0.99	0.19	0.32	802
accuracy			0.55	4800
macro avg	0.74	0.55	0.55	4800
weighted avg	0.74	0.55	0.55	4800

Conclusion

Supervised learning excels with high-quality labeled data, but acquiring it can be costly. Unsupervised learning offers promise, especially with powerful models, by leveraging unlabeled data. This could significantly reduce reliance on manual labeling. Further exploration of unsupervised learning for detection tasks is essential.



Thank You