CSE 424

MITIGATING BIAS IN CYBER-BULLYING DETECTION

SECTION 02

GROUP/ROW NO. 04 MEMBERS

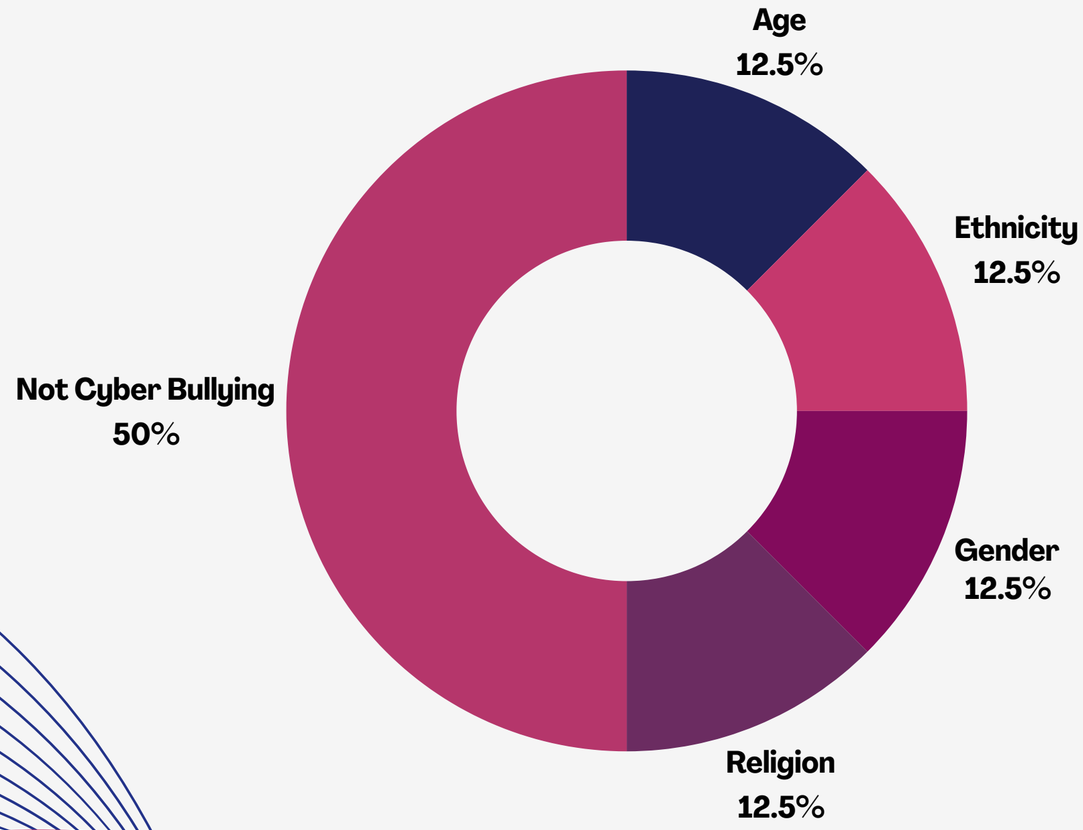
- 21201286 Fahmid Hasan Chowdhury
- 21201357 Zakaria Ibne Rafiq
- 21201434 Maliha Binte Masud
- 23241086 Md. Sabbir Hossain Mirza
- 23241099 Amina Zannat Nurhan
- 23241100 MD Rakibul Hasan Talukder
- 23241118 Abdur Rahman Shafi
- 24141150 Radito Dhali

INTRODUCTION

The rise of social media brings connection, but also negativity like cyberbullying. We propose a system to address context-dependent cyberbullying detection on Twitter. Our system will consider factors like age, gender, and intent behind messages, going beyond just "bad" words. We'll compare three machine learning algorithms to find the best for unbiased and accurate detection. This research contributes to a safer online environment and highlights the impact of data and algorithm selection.

DATASET AND APPROACH

- Leverage publicly available dataset of 48,000 cyberbullying tweets.
- Addresses class imbalance and captures diverse cyberbullying forms (age, gender, etc.).
- Utilizes Dynamic Query Expansion (DQE) to gather more natural data points.
- Provides a balanced multiclass dataset for representative cyberbullying detection.
- Rich data fuels development of a robust cyberbullying detection model.



DATASET AND APPROACH

Balanced Datasets for Cyberbullying Detection

- **Problem:** Imbalanced social media data (more non-cyberbullying examples) can lead to models that miss cyberbullying.
- **Solution:** Balanced datasets with equal cyberbullying & non-cyberbullying texts.
- **Benefits:** Fairer evaluation & improved model generalizability for real-world data.

METHODOLOGY

Dataset

PRE-PROCESSING

Text Cleaning for Consistency

We ensure all text is lowercase and punctuation is removed. This helps the model treat variations like "Happy" and "happy" as the same.

Common words with little meaning, like "the" or "is," are identified and removed using the Natural Language Toolkit (NLTK) library. This reduces noise and lets the model focus on content-rich terms.

Finally, we apply stemming or lemmatization to reduce words to their base form. Stemming often creates a root word (e.g., "running" -> "run"), while lemmatization aims for a grammatically correct base form (e.g., "running" -> "run"). This improves efficiency and captures the semantic similarity between words.

MACHINE LEARNING MODELS

01 Naive Bayes

- Simple and efficient classifier.
- Works well with high-dimensional data (like text).
  - Makes strong independent assumptions between features (words).This might limit its effectiveness for complex, nuanced cyberbullying detection where context matters.
- Fast training and good for initial exploration.

02 Logistic Regression

- Works well with balanced datasets.
- Interpretable model: Understand which words contribute most to cyberbullying classification.
- Efficient training and good for initial exploration.
- May struggle with highly non-linear relationships between features (words) in complex cyberbullying text.

MACHINE LEARNING MODELS

03 K-Nearest Neighbors (KNN)

- Simple and efficient classifier based on local similarity.
- No explicit training phase—classifies new data points based on the labels of its k nearest neighbors in the training data.
- Effective for well-defined data clusters and interpretable results (neighbors reveal similar examples).
- Computational cost increases with data size and may struggle with high-dimensional data (like text).

04 Neural Network

- Recognize patterns in data, resembling brain functionality.
- Composed of interconnected nodes (neurons).
- Process information and learn through adjustments.
- Layered structure: input → hidden layers → output.
- Hidden layers crucial for learning complex patterns.
- Learn via backpropagation: adjust connections based on prediction error.
- Error minimization over training iterations improves performance.

RESULT & COMPETITIVE ANALYSIS

NAIVE BAYES

class	precision	recall	f1-score	support
Age	0.87	0.73	0.79	632
Ethnicity	0.88	0.83	0.85	607
Not Cyberbullying	0.81	0.89	0.85	2294
Religion	0.82	0.89	0.85	588
Gender	0.86	0.64	0.73	612
Accuracy			0.83	4733
Macro avg	0.85	0.79	0.82	4733
Weighted avg	0.83	0.83	0.83	4733

K NEAREST NEIGHBOUR (KNN)

class	precision	recall	f1-score	support
Age	0.73	0.94	0.82	632
Ethnicity	0.84	0.72	0.78	607
Not cyberbullying	0.73	0.87	0.79	2294
Religion	0.93	0.52	0.67	588
Gender	0.93	0.52	0.67	612
accuracy			0.77	4733
macro avg	0.83	0.71	0.75	4733
weighted avg	0.8	0.77	0.76	4733

LOGISTIC REGRESSION

class	precision	recall	f1-score	support
Age	0.94	0.88	0.91	632
Ethnicity	0.98	0.95	0.96	607
Not Cyberbullying	0.87	0.95	0.91	2294
Religion	0.95	0.9	0.92	588
Gender	0.92	0.73	0.81	612
accuracy			0.91	4733
macro avg	0.93	0.88	0.9	4733
weighted avg	0.91	0.91	0.91	4733

NEURAL NETWORK

class	precision	recall	f1-score	support
Age	0.82	0.78	0.8	632
Ethnicity	0.88	0.83	0.86	607
Not Cyberbullying	0.84	0.85	0.84	2294
Religion	0.78	0.83	0.8	588
Gender	0.72	0.7	0.71	612
accuracy			0.82	4733
macro avg	0.81	0.8	0.8	4733
weighted avg	0.82	0.82	0.82	4733

RESULT & ANALYSIS

- Logistic regression excelled in cyberbullying detection on our balanced dataset.
- It achieved high accuracy, precision, recall, and F1-score.

This means:

- Accurately classifies cyberbullying and non-cyberbullying texts.
- Minimizes false positives (wrongly accusing someone).
- Performs well across all classes (balanced metrics).

- Logistic regression is a strong candidate for cyberbullying detection in our scenario.

CONCLUSION

TOWARDS A SAFER ONLINE SPACE

BREAKTHROUGH

- Our project developed a machine learning model for effective cyberbullying detection.
- It incorporates demographics (age, race, gender) for a nuanced understanding.
- Logistic regression outperformed KNN, Naive Bayes, and Neural Networks in our evaluation.

FUTURE WORK

- Sarcasm and Irony Detection: Improve understanding of true intent behind messages.
- Big Data Integration: Increase accuracy with more training data (consider computational cost).
- Unsupervised Learning: Explore new patterns and trends in cyberbullying behavior.

Group/Row No. 04

CSE 424 : PATTERN RECOGNITION

THANK YOU

HTTPS://GITHUB.COM/NOT-SOLUCKY/424PROJECT