Mitigating Bias in Cyberbully Detection from Social Media with Text Classification

Zakaria Ibne Rafiq [21201357], Fahmid Hasan Chowdhury[21201286],
Radito Dhali,[24141150] Rakibul Hasan Talukdar[23241100], Sabbir Hossain Mirza[23241086],
Amina Jannat Nurhan[23241099], Maliha Ishrar[21201434], Abdur Rahman Shafi[23241118]
BRAC University
Dhaka, Bangladesh

***Abstract*—With the rise of social media, cyberbullying has become a prevalent issue, causing significant emotional distress. Our model employs machine learning algorithms to categorize text based on demographic factors, including age, ethnicity, and religious beliefs. This project explores the use of machine learning for cyberbullying detection on Twitter. We train various algorithms such as Logistic Regression, Naive Bayes and K-nearest neighbour on a dataset of cyberbullying tweets. Our analysis reveals that Logistic Regression achieves the highest accuracy in identifying offensive language, highlighting the importance of selecting appropriate algorithms and prerocessing techniques for effective cyberbullying detection.**

***Index Terms*—Twitter Naive Bayes, Logistic Regression, KNN, SVM**

I. Introduction

The rise of social media has brought a wave of connection, instant news, educational resources, and entertainment. However, this digital landscape also fosters negativity, with cyberbullying being a particularly harmful issue.

Researchers are actively tackling cyberbullying, and our project contributes to this effort. We aim to develop a highly accurate cyberbullying detection system that considers factors like age, gender, language, race, and nationality. We train our model on a dataset meticulously collected by Jason Wang using Dynamic Query Expansion (DQE), a semi-supervised technique. He extracted natural Twitter conversations focused on specific cyberbullying categories. To prepare the data for analysis, we employ the NLTK toolkit for Natural Language Processing (NLP) tasks. This includes lemmatization, which converts words to their base form (e.g., ”running” becomes ”run”), and the removal of stop words (common words like ”the” and ”a”) that don't contribute significantly to meaning. Finally, we train three machine learning algorithms: Naive Bayes, Logistic Regression, and K-Nearest Neighbors. By comparing their performance metrics (accuracy, precision, recall, F1-score), we can identify the most effective model for cyberbullying detection.

Our proposed system strives to be an unbiased and valuable tool for detecting cyberbullying text on Twitter. This research not only contributes to creating a more positive online environment but also demonstrates the influence of data preprocessing and machine learning algorithm selection on model performance. Background info goes here.. [1]

References

[1] J. Wang, K. Fu, and C.-T. Lu, “Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection,” in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1699–1708.