LatIA. 2025; 3:126
doi: 10.62486/latia2025126
ISSN: 3046-403X

ORIGINAL

Analysis of Cyberbullying Behaviors Using Machine Learning: A Study on Text Classification

Análisis de los comportamientos de ciberacoso mediante aprendizaje automático: Un estudio sobre clasificación de textos

Alok Kumar Anand[1], Rajesh Kumar Mahto[1], Awadesh Prasad[1]

[1]Department of Computer Science, V. K. S. University, Ara-802301, India.

Cite as: Kumar Anand A, Kumar Mahto R, Prasad A. Analysis of Cyberbullying Behaviors Using Machine Learning:A Study on Text Classification. LatIA. 2025; 3:126. https://doi.org/10.62486/latia2025126

Submitted: 23-03-2024 **Revised:** 11-06-2024 **Accepted:** 18-09-2024 **Published:** 01-01-2025

Editor: Dr. Rubén González Vallejo

ABSTRACT

Introduction: Cyberbullying is a significant concern in today's digital age, affecting individuals across various demographics.
Objective: this study aims to analyze and classify instances of cyberbullying using a dataset sourced from Kaggle, containing text data labeled for different types of bullying behaviors.
Method: our approach to tackling these challenges involves several key steps, starting with data preprocessing and feature extraction to identify patterns and improve detection methods, enhancing our understanding of how cyberbullying manifests in online communications.
Result: the dataset provides a valuable resource for developing and evaluating machine learning models aimed at detecting sexist and racist content in tweets.
Conclusion: this study advances the current understanding of the complexities involved in detecting cyberbullying and paves the way for future breakthroughs in this domain. The binary classification enabled by the 'oh_label' column streamlines the analysis process, making it particularly compatible with binary classification models.

Keywords: Security; Privacy; Cyberbullying; Dark Web.

Introducción: el ciberacoso es una preocupación importante en la era digital actual, que afecta a individuos de diversos grupos demográficos.
Objetivo: este estudio tiene como objetivo analizar y clasificar los casos de acoso cibernético utilizando un conjunto de datos procedentes de Kaggle, que contiene datos de texto etiquetados para diferentes tipos de comportamientos de acoso.
Método: nuestro enfoque para abordar estos desafíos implica varios pasos clave, comenzando con el preprocesamiento de datos y la extracción de características para identificar patrones y mejorar los métodos de detección, mejorando nuestra comprensión de cómo se manifiesta el ciberacoso en las comunicaciones en línea.
Resultados: el conjunto de datos proporciona un valioso recurso para desarrollar y evaluar modelos de aprendizaje automático destinados a detectar contenidos sexistas y racistas en tuits.
Conclusiones: este estudio avanza en la comprensión actual de las complejidades implicadas en la detección del ciberacoso y allana el camino para futuros avances en este dominio. La clasificación binaria que permite la columna 'oh_label' agiliza el proceso de análisis, haciéndolo especialmente compatible con modelos de clasificación binaria.

Palabras clave: Seguridad; Privacidad; Ciberacoso; Dark Web.

INTRODUCTION

The widespread adoption of social media and digital communication platforms has transformed the way people interact and share information. However, this connectivity has also brought about significant challenges, including the pervasive issue of cyberbullying. Cyberbullying refers to the use of electronic communication to harass, threaten, or demean others, often with anonymity that emboldens perpetrators.(1) This harmful behavior is not limited by age, gender, or location and can have profound psychological impacts on victims. Individuals who experience cyberbullying often report heightened levels of stress, anxiety, and depression, and in severe cases, it can contribute to suicidal ideation and self-harm. The digital nature of cyberbullying means that harmful messages can spread rapidly and are accessible around the clock, making the harassment feel relentless and inescapable.(2,3)

Given the serious implications of cyberbullying, there is an urgent need for effective detection and intervention strategies. One promising approach is the development of automated systems that can accurately identify instances of cyberbullying in textual data. Such systems can play a crucial role in moderating online content and providing timely alerts that allow for appropriate interventions. Machine learning offers powerful tools for building these automated systems, as it enables the classification and analysis of large volumes of data with a high degree of accuracy.(4) By leveraging machine learning, we can train models to recognize patterns and features characteristic of cyberbullying, even when they are expressed in diverse and subtle ways.

Cyberbullying is a pervasive issue that affects individuals of all ages, but it poses a particular threat to adolescents, who are especially susceptible to its harmful impacts. Recent research, such as the study conducted by Kowalski et al.(5) (2023), suggests that between 20 % and 40 % of teenagers have encountered cyberbullying at some stage, with prevalence rates influenced by factors including gender, ethnicity, and socio-economic background. Findings indicate that girls are more frequently targeted than boys, often subjected to harassment that focuses on their appearance or social standing.

Cyberbullying manifests in a variety of forms, including overt behaviors like insults and threats, as well as covert actions such as spreading false information or sharing private details without permission. The anonymity afforded by digital platforms can amplify the intensity of cyberbullying, as offenders often feel shielded from direct accountability (Barlett et al., 2024). Moreover, the public and enduring nature of online content means that harmful posts can be disseminated widely and persist indefinitely, leading to prolonged exposure to bullying and extended psychological distress for the victims.(6)

Addressing cyberbullying effectively requires not only detection but also proactive intervention and prevention strategies. Contemporary research highlights the need for a comprehensive approach that integrates educational initiatives, policy reforms, and technological innovations.(7)

Educational programs play a vital role in combating cyberbullying by increasing awareness among students, parents, and educators about its impact and encouraging responsible online behavior. Williford et al. (2023) investigated school-based interventions that focus on teaching digital citizenship and fostering empathy. Their findings indicate that such initiatives can significantly lower the incidence of cyberbullying by promoting a culture of respect and positive interactions online.(8)

Objective

This research focuses on utilizing a dataset from Kaggle that includes text data labeled into six distinct categories: Age, Ethnicity, Gender, Religion, Other, and Not Cyberbullying. Each category represents a different form of bullying or non-bullying communication, allowing for a nuanced analysis of how cyberbullying manifests across various contexts. The dataset provides a valuable resource for training machine learning models to differentiate between harmful and benign content. Our objective is to develop a robust model capable of classifying text into these categories with high precision, thereby enhancing the ability of automated moderation systems to identify and respond to cyberbullying in real-time.

Challenges in Cyberbullying Detection

Detecting cyberbullying in textual data is inherently challenging due to the variability and complexity of human language. Cyberbullying can be overt, involving explicit threats or insults, or it can be more subtle, such as through sarcasm, exclusion, or indirect commentary. Additionally, language evolves rapidly, and slang, memes, or coded language can complicate detection efforts. The same word or phrase can have different connotations depending on the context, making it difficult for basic keyword-based systems to accurately identify harmful content. Therefore, any effective detection system must account for these nuances and be adaptable to the dynamic nature of online communication. Another challenge lies in balancing sensitivity and specificity.(9,10) A model that is too aggressive in flagging content may produce a high rate of false positives, leading to unnecessary censorship and frustration among users. Conversely, a model that is too lenient may fail to identify harmful content, allowing bullying behavior to persist. Striking the right balance requires careful calibration of the model and continuous refinement based on feedback and new data.

METHOD

Our approach to tackling these challenges involves several key steps, starting with data preprocessing and feature extraction. Preprocessing is essential for cleaning the text data and standardizing it in a way that is suitable for analysis. This step includes removing punctuation, converting text to lowercase, eliminating stop words, and applying stemming and lemmatization to reduce words to their root forms. These actions help to simplify the text data and focus on the core components that contribute to meaning. For feature extraction, we utilize the Term Frequency-Inverse Document Frequency (TF-IDF) method, which converts the processed text into numerical vectors that reflect the importance of each term relative to the entire dataset.(11,12) TF-IDF helps to highlight words that are more indicative of bullying behavior, thus providing the machine learning model with the features needed to make accurate classifications.

Cyberbullying Data Analysis

This dataset is a collection of datasets from different sources related to the automatic detection of cyber-bullying. The data is from different social media platforms like Kaggle, Twitter, Wikipedia Talk pages and YouTube. The data contain text and labeled as bullying or not. The data contains different types of cyber-bullying like hate speech, aggression, insults and toxicity.(13)

The dataset consists of 16 851 entries with five columns: index, id, Text, Annotation, and oh_label. Here's a breakdown of the dataset:

Data Overview

- The dataset contains tweets with annotations indicating whether the tweet is related to sexism, racism, or none.
- The oh_label column is a binary indicator where 1 represents the presence of sexism or racism, and 0 represents none.

Summary Statistics

- The oh_label column has a mean of approximately 0,317, indicating that around 31,7 % of the tweets are labeled as either sexism or racism.
- The distribution of oh_label is skewed towards 0, as seen in the summary statistics.

Annotation Distribution

- The Annotation column has three unique values: none, sexism, and racism.
- The majority of the tweets (11,501) are labeled as none, followed by sexism (3 377) and racism (1 970).

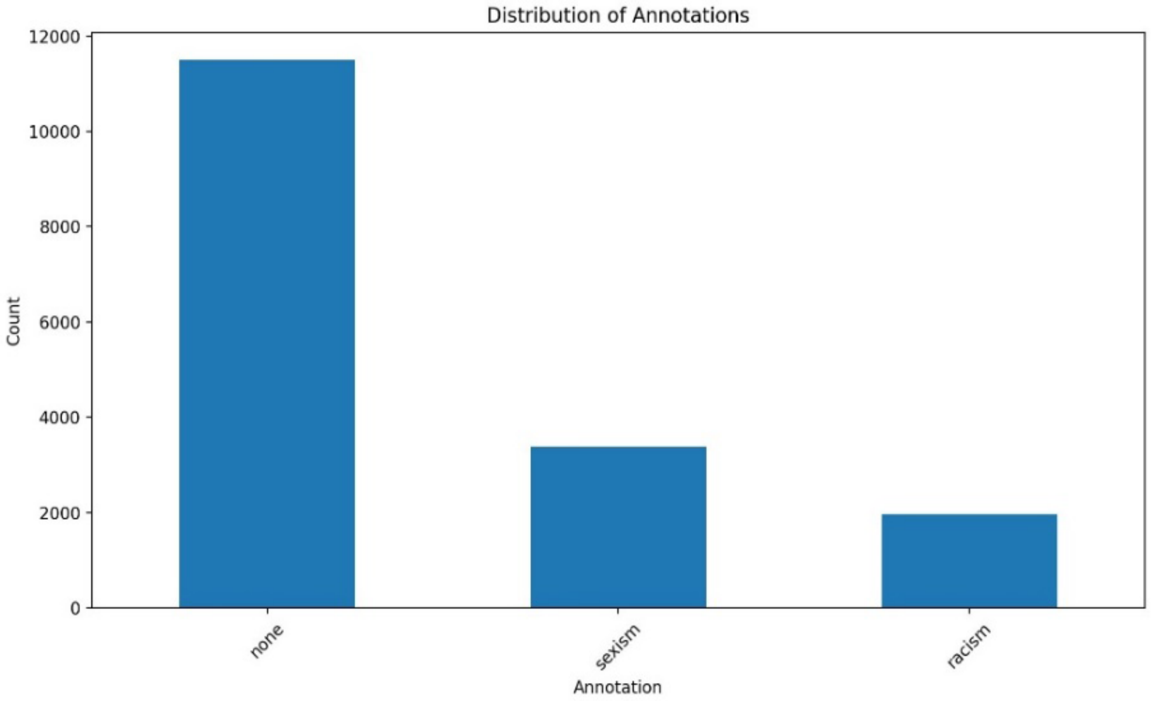


Figure 1. The bar plot illustrates the distribution of annotations in the dataset

This dataset is useful for analyzing the prevalence of sexism and racism in tweets. The annotations provide a clear categorization, allowing for targeted analysis and potential machine learning applications to automatically detect such content in social media posts. The imbalance in the dataset, with a higher number of none annotations, suggests that any model trained on this data should account for this imbalance to avoid bias towards the majority class.

Pre-processing

Pre-processing is a crucial step in preparing the dataset for analysis or machine learning tasks. It involves cleaning and transforming raw data into a format that is more suitable for analysis. Here's a detailed explanation of the pre-processing steps that can be applied to the dataset:

Data Cleaning:

- Handling Missing Values: Check for any missing values in the dataset. In this case, the dataset appears to have no missing values in the columns, as indicated by the non-null counts in the data overview.
- Removing Duplicates: Identify and remove any duplicate rows to ensure that each entry is unique. This can be done using the drop_duplicates() function in pandas.

Text Pre-processing:

- Tokenization: Split the text in the Text column into individual words or tokens. This is a fundamental step in natural language processing (NLP) to analyze the text data.
- Lowercasing: Convert all text to lowercase to ensure uniformity, as text data can be case-sensitive.
- Removing Punctuation and Special Characters: Strip out punctuation and special characters from the text to focus on the words themselves.
- Stop Words Removal: Remove common stop words (e.g., "and", "the", "is") that do not contribute much to the meaning of the text.
- Stemming/Lemmatization: Reduce words to their base or root form. Stemming cuts off prefixes or suffixes, while lemmatization uses vocabulary and morphological analysis.

Encoding Categorical Variables:

- The Annotation column contains categorical data with values like none, sexism, and racism. These can be encoded into numerical values using techniques like one-hot encoding or label encoding. This is essential for machine learning models that require numerical input.

Feature Engineering:

- Creating New Features: Derive new features from the existing data. For example, the length of the tweet or the number of hashtags could be useful features.
- Normalization/Standardization: Scale numerical features to a standard range. This is particularly important if the dataset includes features with different units or scales.

Balancing the Dataset:

- The dataset is imbalanced, with a higher number of none annotations compared to sexism and racism. Techniques like oversampling, undersampling, or using synthetic data generation methods (e.g., SMOTE) can be applied to balance the dataset.

Splitting the Dataset:

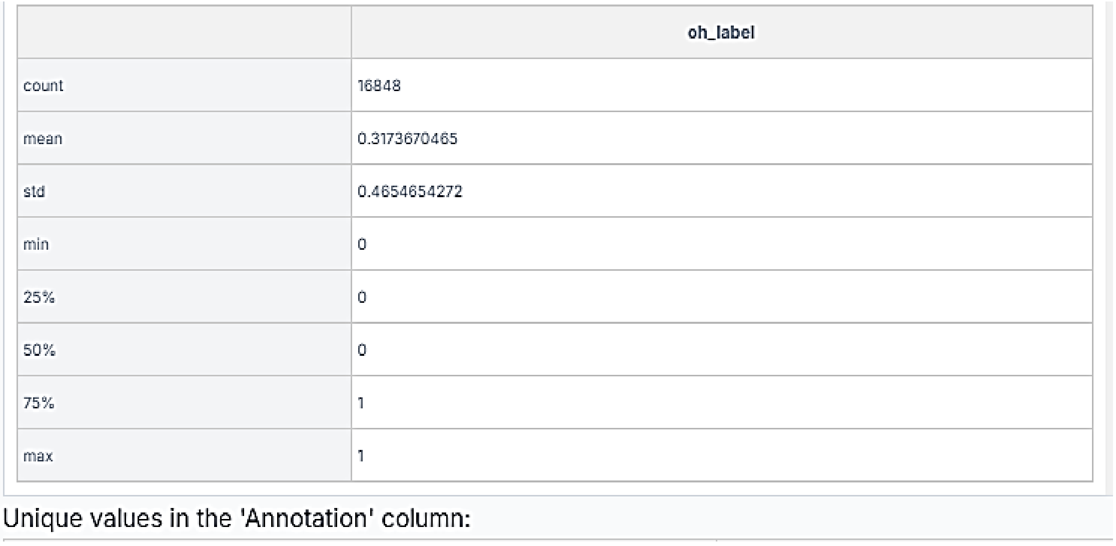
- Divide the dataset into training, validation, and test sets. This is crucial for evaluating the performance of machine learning models and ensuring they generalize well to unseen data.

Data Augmentation:

- For text data, augmentation techniques like synonym replacement or back-translation can be used to increase the diversity of the training data.

By applying these pre-processing steps, the dataset becomes more structured and ready for analysis or model training. Proper pre-processing can significantly enhance the performance of machine learning models and lead to more accurate insights from the data.

Annotation Process

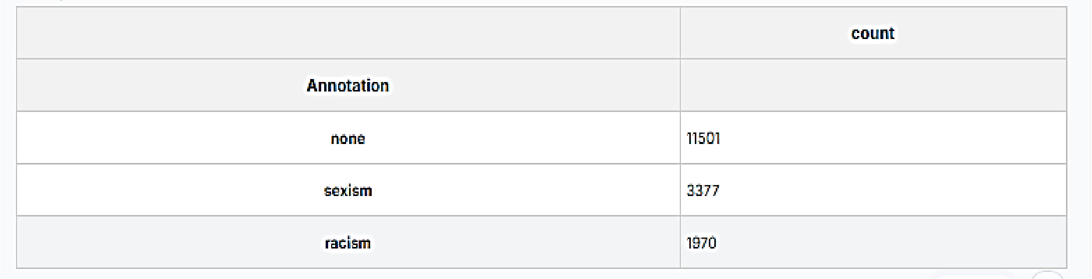


	oh_label
count	16848
mean	0.3173670465
std	0.4654654272
min	0
25%	0
50%	0
75%	1
max	1

Unique values in the 'Annotation' column:

Figure 2. This binary classification (0 or 1) indicates whether a tweet contains offensive content (sexism or racism) or not

The binary classification of tweets into offensive (1) and non-offensive (0) categories serves as a foundational aspect of the dataset, enabling targeted analysis and the development of automated systems to identify and address offensive content on social media platforms. This classification approach is essential for understanding the prevalence of such content and for implementing measures to mitigate its impact.



	count
Annotation	
none	11501
sexism	3377
racism	1970

Figure 3. Annotation

The manual annotation process is essential for creating a reliable dataset that can be used for various applications, including sentiment analysis, content moderation, and the development of algorithms aimed at detecting offensive language in social media.

Implications and Future Directions

The results of this study underscore the potential of machine learning in enhancing cyberbullying detection and prevention efforts. Automated systems equipped with well-trained models can significantly aid content moderation teams by providing preliminary assessments of text, allowing human moderators to focus on the most critical cases. However, the nuanced nature of language and the evolving landscape of online communication mean that continuous updates and improvements to these models are necessary.(14,15) Future research could explore the integration of advanced natural language processing techniques, such as transformers and BERT (Bidirectional Encoder Representations from Transformers), which have shown superior performance in other text classification tasks. Additionally, expanding the dataset to include more diverse examples of cyberbullying, including multimedia data like images and videos, could provide a more comprehensive approach to detecting harmful behavior in digital spaces.

CONCLUSION

The fight against cyberbullying is a complex and ongoing challenge, but advances in machine learning offer promising tools for mitigating its impact. By developing and refining models that can accurately classify bullying content, we can create safer online environments and support efforts to protect vulnerable individuals from harm. This study represents a step forward in understanding the intricacies of cyberbullying detection and lays the groundwork for future innovations in the field. The dataset provides a valuable resource for developing and evaluating machine learning models aimed at detecting sexist and racist content in tweets.

However, the imbalance in the annotation distribution poses a challenge that must be addressed to avoid bias in model predictions. Techniques such as resampling or using class weights can be employed to mitigate this issue. The binary classification provided by the oh_label column simplifies the task, making it suitable for binary classification models. Overall, this dataset is well-suited for research and development in the field of automated content moderation on social media platforms.

REFERENCES

1. Kowalski R. Cyberbullying. In: The Routledge international handbook of human aggression. Routledge; 2018. p. 131–42.

2. Dhamodharan M, Sunaina K. Cyberbullying: A disturbed psyche and digital abuse in 21st century. In: Analyzing new forms of social disorders in modern virtual environments. IGI Global; 2023. p. 224–49.

3. Haque MA, Ahmad S, Haque S, Kumar K, Mishra K, Mishra BK. Analyzing University Students' Awareness of Cybersecurity. In: 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC). IEEE; 2023. p. 250-7.

4. Haque MA, Ahmad S, Sonal D, Abdeljaber HAM, Mishra BK, Eljialy AEM, et al. Achieving Organizational Effectiveness through Machine Learning Based Approaches for Malware Analysis and Detection. Data Metadata. 2023;2:139.

5. Kowalski RM, Giumetti GW, Feinn RS. Is cyberbullying an extension of traditional bullying or a unique phenomenon? A longitudinal investigation among college students. Int J bullying Prev. 2023;5(3):227–44.

6. Barlett CP. Cyberbullying process in US adolescents and their parents: Testing and extending the Barlett Gentile cyberbullying model. Aggress Behav. 2024;50(1):e22117.

7. Haque MA, Ahmad S, Abboud AJ, Hossain MA, Kumar K, Haque S, et al. 6G wireless Communication Networks: Challenges and Potential Solution. Int J Bus Data Commun Netw. 2024;19(1):1–27.

8. Williford A, Depaolis KJ. Predictors of cyberbullying intervention among elementary school staff: The moderating effect of staff status. Psychol Sch. 2016;53(10):1032–44.

9. Hossain MA, Haque MA, Ahmad S, Abdeljaber HAM, Eljialy AEM, Alanazi A, et al. AI-enabled approach for enhancing obfuscated malware detection: a hybrid ensemble learning with combined feature selection techniques. Int J Syst Assur Eng Manag [Internet]. 2024; Available from: https://doi.org/10.1007/s13198-024-02294-y

10. Haque MA, Ahmad S, John A, Mishra K, Mishra BK, Kumar K, et al. Cybersecurity in Universities: An Evaluation Model. SN Comput Sci [Internet]. 2023;4(5):569. Available from: https://doi.org/10.1007/s42979-023-01984-x

11. Sabella RA, Patchin JW, Hinduja S. Cyberbullying myths and realities. Comput Human Behav. 2013;29(6):2703–11.

12. Olweus D. Cyberbullying: An overrated phenomenon? Eur J Dev Psychol. 2012;9(5):520–38.

13. Cyberbullying Dataset [Internet]. Available from: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset

14. Haque MA, Sonal D, Haque S, Kumar K, Rahman M. The Role of Internet of Things (IoT) to Fight against Covid-19. In: Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence [Internet]. New York, NY, USA: ACM; 2021. p. 140–6. Available from: https://dl.acm.org/doi/10.1145/3484824.3484900

15. Haque MA, Haque S, Zeba S, Kumar K, Ahmad S, Rahman M, et al. Sustainable and efficient E-learning internet of things system through blockchain technology. E-Learning Digit Media [Internet]. 2023;0(0):1-20. Available from: https://journals.sagepub.com/doi/abs/10.1177/20427530231156711

AVAILABILITY OF DATA AND MATERIALS

The datasets used in this research are publicly available(Kaggel) and properly cited in our dataset section for transparency and ease of replication.

COMPETING INTERESTS SECTION

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Conceptualization: Alok Kumar Anand, Rajesh Kumar Mahto and Awadesh Prasad.
Investigation: Alok Kumar Anand, Rajesh Kumar Mahto and Awadesh Prasad.
Methodology: Alok Kumar Anand, Rajesh Kumar Mahto and Awadesh Prasad.
Writing - original draft: Alok Kumar Anand, Rajesh Kumar Mahto and Awadesh Prasad.
Writing - review and editing: Alok Kumar Anand, Rajesh Kumar Mahto and Awadesh Prasad.