IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 2, February 2025

# Hate Speech and Offensive Content Detection Using Machine Learning

**Mr. Mihir Parekh[1] and Dr. Pallavi Devendra Tawde[2]**
Student, Department of MSc. IT[1]
Assistant Professor, Department of IT & CS[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India
mihirparekh2003@gmail.com and pallavi.tawde09@gmail.com

**Abstract:** *Social media offers global connectivity but also facilitates the spread of hate speeches and lacks offensive language, creating tremendous challenges. Since user-generated content is voluminous, manual moderation using the input is virtually impossible, hence machines' learning (ML) solutions are conceivably essential. The present work assesses five ML models for automatically detecting hate speech, classifying tweets into three classes, namely hate speech, offense language, or neutral speech. Out of 24,783 tweets, XGBoost achieved the highest accuracy, became the best model. For interpretability, feature importance, confusion matrices, and visualization techniques such as word clouds and tweet-length distributions were investigated. While ML models effectively classify the texts, detecting implicit hate speech and multilingual content remains a hurdle. Future work should seek to investigate better models and contextual analysis for safer spaces for interactions on the Internet.*

**Keywords:** Hate Speech Detection, Machine Learning, Text Classification, XGBoost, Naive Bayes, Data Visualization, Social Media Analysis, Automated Content Moderation.

DOI: 10.48175/IJARSCT-23344