JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Offensive Hate Speech Detection Using Machine Learning: A Review

**Mrs.Anita S. Biradar[1], Dr. Aparna S. Hambarde [2,]**
[1]*PG Student*
[2]*Faculty*
*KJ College of Engineering & Management Research, Pune*

***Abstract***— Offensive hate speech is a crime that has become increasingly common, both online and offline. This is due to several variables. On the one hand, the anonymity provided by the internet, particularly social networks, increases the likelihood of aggressive behavior.On the other hand, hate speech has proliferated as a result of people's increased desire to voice their thoughts online. Governments and social media corporations can both benefit from detection and prevention methods because of the harm that this type of discriminatory speech can cause to society. We address this dilemma by offering a comprehensive assessment of the field's research through this survey.

***Keywords***– Offensive Hate Speech, Social Media text, Hate Speech Classification, Natural Language Processing.

## I. INTRODUCTION

The prevalence of offensive hate speech has increased in recent years, both in person and online. Numerous things are involved in this situation People are more likely to act hostilely because of the anonymity of the internet, but they are also more likely to voice their thoughts online, which helps hate speech proliferate [1]. Because discriminatory speech can have a disastrous impact on society, governments and social media companies can benefit from detection and prevention strategies.

We hope that our survey may shed some light on the numerous studies that have been carried out in this field[2][3].

Offensive Hate speech is defined as any discourse that has the capacity to do harm to an individual or group and that may result in violence, insensitivity, or illogical or inhuman behavior. The prevalence of hate speech on online social media sites like Facebook and Twitter has increased along with their popularity. There is evidence that hate speech is contributing to an increase in hate crimes[7]. Many government-led initiatives, including the Council of Europe's No Hate Speech campaign, are being implemented as the problem of hate speech gathers pace. Another method it has been implemented is through the EU Hate Speech Code of Conduct, which all social media platforms must sign and abide by within 24 hours[5].

Numerous issues that have been brought to light have generated serious concerns about dataset quality, which is what this study attempts to solve. This work also tackles the second problem, which is that before creating an appropriate classifier, the best features for identifying hate speech must be researched and identified. The most prevalent categories, according to FBI hate crime data, are race, ethnicity, and religion. Datasets typically fit into one of these categories as a result[11].

## II. LITERATURE SURVEY

**ABIDA KOUSAR et al.[2024]** This paper introduces MLHS-CGCapNet, a multilingual hate speech detection model based on convolutional and bidirectional gated recurrent units, as well as a capsule network. Our suggested model performs well on an unbalanced dataset, with accuracy, recall, and f-scores of 0.89, 0.80, and 0.84, respectively. Despite having imbalanced data, the suggested model performs well during training and validation, with accuracy values of 0.93 and 0.90, respectively. Our model beats both the baseline and cutting-edge approaches.

**Patrick Burkart et al.[2021]**The 2017 "Unite the Right" rally in Charlottesville, Virginia, was a meticulously planned multi-dimensional event that sought to unify white nationalist and neo-Nazi groups through highly symbolic, aggressive, and terrorizing means. Using symbolic interactionism and levels of analysis theory, this study examines how the event unfolded across three distinct "theatres": symbolic (rituals, symbols, and propaganda), systems-technical (strategic organization and digital influence),

**MPH, Xiang Xu, et al. [2021]** this study examined the link between the terms "COVID-19" and "Chinese virus" on Twitter and anti-Asian sentiment. Analyzing over 1.2 million hashtags, it found that 50.4% of posts with #chinesevirus showed anti-Asian sentiment, compared to 19.7% with #covid19. Anti-Asian sentiment surged significantly following President Trump's use of "Chinese Virus." The findings support using "COVID-19" as a less-stigmatizing term.

**Tomer Wullach et al. [2023]** this study addresses limitations in hate speech detection by creating a dataset of 1 million realistic hate and non-hate sequences using a deep generative language model. Training a DL-based detector on this expanded dataset significantly improves detection performance across multiple public datasets, advancing toward a more robust and fully automated hate speech detection solution.

**JAMEEL AHMAD et al.[2024]** This study tackles the challenge of multilingual hate speech detection across 12 languages by developing MLHS-CGCapNet, a lightweight model combining convolutional and gated recurrent units with a capsule network. Achieving high accuracy (0.93 training, 0.90 validation), the model outperforms existing techniques, especially on imbalanced datasets.

**ZAINAB MANSUR et al.[2023].**This review analyzes 91 studies on Twitter hate speech detection from 2015 to 2022, summarizing current methods and identifying gaps for improvement. It aims to guide researchers in developing more effective models, ultimately aiding social platforms in preemptively identifying and reducing hate speech.

**Lingyao Li et al.[2023]** This study compares ChatGPT's ability to detect harmful content (hateful, offensive, and toxic) on social media against human annotations from MTurkers. ChatGPT achieves approximately 80% accuracy, with more consistent classification of non-HOT comments. The study emphasizes the importance of prompt options on performance and demonstrates ChatGPT's ability to moderate enormous amounts of content.

**Shyamal Mishra et al.[2024]**This study explores using ChatGPT, a zero-shot LLM, to detect toxicity in GitHub discussions, addressing the challenge of limited and imbalanced software engineering toxicity datasets. Preliminary results show ChatGPT's potential, with further investigation needed for improving interpretability and integration into developer communication channels.

**Xinlei He et al. [2024]** This study explores using large language models (LLMs) and prompt learning to detect and mitigate toxic content online. The results show that LLMs outperform traditional models in toxicity classification, toxic span detection, and detoxification, with a 10% improvement in

classification accuracy and successful toxicity reduction while maintaining semantic meaning.

**Dimple Patil et al.[2024].**This chapter examines the challenges of implementing generative AI in businesses, including data privacy concerns, model biases, scalability issues, and high computational costs. It offers actionable insights to help companies navigate these hurdles for sustainable and scalable AI adoption.

**MUHAMMAD BILAL et al.[2022]** This work creates a Roman Urdu Hate Speech Detection model utilizing Bi-LSTM and an attention layer, which achieves high performance with 87.5% accuracy and 88.5% F1 score. The model outperforms classical machine learning and other deep learning models, and lexical normalization improves its performance even further.

**CHIH-CHIEN WANG et al.[2022]**This study develops a Chinese hate speech lexicon and a detection system using both deep learning and lexicon-based approaches. It contributes a dataset and lexicon for detecting political hate speech, offering methods that can be applied to other languages as well.

**AYESHA ATIF et al.[2024]** This study develops a cyberbullying detection model for Roman Urdu tweets using GRU and word2vec, achieving 97% accuracy and F1-measure. It also introduces a temporal analysis method for bullying behavior, outperforming other machine learning models.

**K. SREELAKSHMI et al.[2024]** This study looks upon multilingual transformer models for identifying hate speech and foul language in CodeMix Dravidian languages. The MuRIL model with SVM obtained great accuracy (96% for Malayalam, 72% for Tamil, and 66% for Kannada), while a cost-sensitive learning method enhanced class balance. A unique annotated Malayalam-English dataset was also presented.

**GIL RAMOS et al.[2024]** This research focuses on detecting hate speech in European Portuguese using transfer learning models. BERTimbau was the highest performing model for YouTube comments, with an F-score of 87.1%, while GPT-3.5 performed best for Twitter, with an F-score of 50.2%. The study also emphasizes the importance of in-domain training and contextualization in generative model prompts.

**JUAN MANUEL PÉREZ et al.[2023]**This study explores the impact of adding contextual information to hate speech detection on Twitter, specifically in replies to digital newspaper posts about COVID-19. Using a Spanish corpus, the study finds that incorporating context improves performance, increasing Macro F1 by 4.2 and 5.5 points for binary and multi-label tasks, respectively.

## III. **OPEN ISSUES:-**

• No survey on automatic detection, hate speech detection in Tex was disregarded in earlier technology.
• Compared to "hateful" sentences, there are far more sentences in the White Supremacy Forum that do not express hate speech.

• The number of 'Followers' may have been the individual feature that contributed to the increase in the F1-score on the two datasets, which also enhanced the subset improvement.
• For upcoming research initiatives, these unigrams and patterns can be utilized as pre-existing dictionaries that are not part of the suggested hate speech detection dictionaries.

## CONCLUSION

The multi-class automated hate speech categorization for text problem is resolved with noticeably better outcomes after the main issues have been identified. Hate speech can be categorized into 10 different binary datasets. A group of professionals annotated each dataset, strictly adhering to a set of rules. Every piece of information was dispersed equally throughout the various datasets. Their ability to use subtlety in language was also improved. Such a dataset had to be generated to close the gap in the field.

## REFERENCES


[1] ABIDA KOUSAR , JAMEEL AHMAD 2 , KHALID IJAZ , AMR YOUSEF , ZAFFAR AHMED SHAIKH,IKRAMULLAH KHOSA, DURGA CHAVALI, AND MOHD ANJUM" MLHS-CGCapNet: A Lightweight Model for Multilingual Hate Speech Detection" IEEE ACCESS 2024

[2] Emily Blouta and Patrick Burkartb1.,"White Supremacist Terrorism in Charlottesville: Reconstructing 'Unite the Right'"2021

[3] Yulin Hswen, ScD, MPH, Xiang Xu.,"Association of "#covid19" Versus "#chinesevirus" With Anti-Asian Sentiments on Twitter: March 9–23, 2020".,2021

[4] Tomer Wullach.,"Towards Hate Speech Detection at Large via Deep Generative Modeling",2023

[5] ABIDA KOUSAR 1 , JAMEEL AHMAD 2 , KHALID IJAZ 1 , AMR YOUSEF 3,4, (Member, IEEE), ZAFFAR AHMED SHAIKH 5,6, (Member, IEEE), IKRAMULLAH KHOSA 7 , (Senior Member, IEEE).,"MLHS-CGCapNet: A Lightweight Model for Multilingual Hate Speech Detection".,2024

[6] ZAINAB MANSUR 1,2, NAZLIA OMAR 1 , AND SABRINA TIUN 1.,"Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities".,2023

[7] Lingyao Li1 *, Lizhou Fan1 , Shubham Atreja1 , Libby Hemphill1.,""HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media",2023

[8] Shyamal Mishra Drexel university Philadelphia, Pennsylvania, USA sm4825@drexel.edu Preetha Chatterjee.,"Exploring ChatGPT for Toxicity Detection in GitHub",2024

[9] Xinlei He1 Savvas Zannettou2 Yun Shen3 Yang Zhang1.,"You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content"2024

[10] Dimple Patil 1 , Nitin Liladhar Rane 2 , Jayesh Rane 3.,"Challenges in implementing ChatGPT and generative artificial intelligence in various business sectors".,2024

[11] MUHAMMAD BILAL 1 , ATIF KHAN 1 , SALMAN JAN 2,3, AND SHAHRULNIZA MUSA.,"Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform"2022

[12] CHIH-CHIEN WANG , MIN-YUH DAY, (Member, IEEE), AND CHUN-LIAN WU.,"Political Hate Speech Detection and Lexicon Building: A Study in Taiwan".,2022

[13] AYESHA ATIF1 , AMNA ZAFAR 1 , MUHAMMAD WASIM 2 , TALHA WAHEED 1 , AMJAD ALI 3 , HAZRAT ALI 4 , (Senior Member, IEEE), AND ZUBAIR SHAH3.,"Cyberbullying Detection and Abuser Profile Identification on Social Media for Roman Urdu"2024

[14] K. SREELAKSHMI 1 , B. PREMJITH 1 , BHARATHI RAJA CHAKRAVARTHI 2 , AND K. P. SOMAN1.,"Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach"2024

[15] GIL RAMOS 1 , FERNANDO BATISTA 2,3, (Senior Member, IEEE), RICARDO RIBEIRO 2,3, PEDRO FIALHO 1 , SÉRGIO MORO 1,4,

ANTÓNIO FONSECA 1 , RITA GUERRA 5 , PAULA CARVALHO 3 , CATARINA MARQUES 6 , AND CLÁUDIA SILVA 7,8.,"Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts"2024

[16] JUAN MANUEL PÉREZ 1 , FRANCO M. LUQUE 2,3, DEMIAN ZAYAT 4 , MARTÍN KONDRATZKY 5 , AGUSTÍN MORO3,6, PABLO SANTIAGO SERRATI 3,7 , JOAQUÍN ZAJAC 3,8, PAULA MIGUEL 3,7, NATALIA DEBANDI 9 , AGUSTÍN GRAVANO 3,10,11 , AND VIVIANA COTIK1,12.,"Assessing the Impact of Contextual Information in Hate Speech Detection",2023

IV.