IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 10, October 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310102|

A Machine Learning Framework for Detecting Cyberbullying in Social Networks

Pradeep M, Harshith R, G M Supreeth

Assistant Professor, Department of Information Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

UG Student, Department of Information Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

UG Student, Department of Information Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

ABSTRACT: A major issue of our time-a result of easy access to a digital platform-is cyberbullying, which is considered an urgent global issue of mental health and safety concerns in the world. This article introduces a bilingual cyberbullying detection system that supports both English and Hindi texts, transcribed into the English language. This presents a much-needed multilingual tool in online safety applications. The proposed model is based on the machine learning approach that utilizes text preprocessing, feature extraction through Term Frequency-Inverse Document Frequency (TF-IDF) and classification with Logistic Regression. The pipeline identifies and classifies offensive content with good accuracy while ensuring its real-time detection in a friendly web application developed using Streamlit.

KEYWORDS: Cyberbullying Detection, Machine Learning, Streamlit

I. INTRODUCTION

With the rapid growth of social media around the world, cyberbullying has emerged as a highly prevalent issue with serious implications for mental health and digital well-being. Most existing systems for detecting cyberbullying are designed for a single language, mainly English, which limits their applicability in multilingual contexts. Previous studies in cyberbullying detection focused on machine learning and deep learning approaches such as SVM, Random Forest, and CNN for image and text-based detection(Cyber_Bullying_and_Toxi…). The approaches, however miss the complexity of multilingual environments where cyberbullying cuts across different languages and dialects, often intermingled within the same digital space.

The bilingual cyberbullying detection model, in this research work, deals with both the English and Hindi languages written in transliteration to the English alphabet. This is a unique contribution because the same system implements Term Frequency-Inverse Document Frequency for feature extraction and Logistic Regression for classification, offering real-time efficient and user-friendly detection of cyberbullying on social media platforms. Designed as an interactive web application using Streamlit, it enables seamless detection in both languages, thus increasing accessibility and safety for a broader user base. It addresses the crucial gap in that it not only enhances the detection of non-English content but also throws light on the linguistic and cultural nuances in cyberbullying behavior. This study adds value to the body of knowledge because it shows that multilingual machine learning applications are possible and have effects in real-time online safety, thereby promoting a more inclusive and safer digital ecosystem.

II. LITERATURE SURVEY

As more incidents of cyberbullying continue to happen on social media, research into the automation of detection techniques based on machine learning and NLP is quite intense. The existing research is mainly about the unimodal model-based or single-language detection system; however, the space available is multilingual and multimodal. For example, one solution uses text-based cyberbullying detection using SVM as well as Random Forest classification algorithms with promising results using TF-IDF and word embeddings(Cyber_Bullying_and_Toxi…). Another incorporated SVM with the deep model MobileNetV2 was used for image-based, which showed high classification accuracies on text and also on image data but did not support multi-lingual(Cyber_Bullying_and_Toxi…).

Generally, deep learning models such as CNN and Long Short-Term Memory (LSTM) have improved the accuracy of detecting toxic language, which demands massive datasets and computationally costly resources to train and perform

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310102|

well. In a study using these models, it reflects the issues of dataset availability and language diversity, where most systems fail to generalize across languages due to limited access to high-quality multilingual data (Cyber_Bullying_and_Toxi…).

This work addresses a unique challenge in developing a bilingual cyberbullying detection system that supports the use of both English and Hindi languages. The study applied TF-IDF for feature extraction and Logistic Regression for classification, thereby achieving a real-time detection result with a lightweight accessible architecture suitable for limited resource languages. The bilingual model shows the feasibility of using language-agnostic preprocessing methods in detecting offense in both English and Hindi (transliterated), thereby bridging the linguistic gap in cyberbullying detection systems. This research also contributes a novel application of regular expressions in a multilingual context for the identification of offensive words, even when transliterated. The use of Streamlit for a web interface also enhances real-world usability, making this system adaptable for deployment on various social media platforms.

Overall, this literature survey highlights how the current bilingual approach expands the landscape of cyberbullying detection, providing a significant advancement in safeguarding multilingual digital communities. This review integrates the findings of relevant studies for contribution in solving multilingual cyberbullying and the promotion of easier access to real-time detection.an in the Middle (MITM) developed and new phishing attacks have managed to be a big problem for cybersecurity. First taking the form of misleading emails and “spoofed” websites, today's phishing scams are far more sophisticatedexploiting social engineering to exploit human nature, in order for an unsuspecting target to part with your personal information. In this digital age, studies have revealed attackers are now more selective and targeted methods — such as spear phishing (that targets a specific individual or organization) means higher chances of success for threat actors. In addition, as artificial intelligence has been incorporated to scams and other phishing attacks take advantage of these new capabilities, the message delivers are more realistic than ever before which means companies must count with solid cybersecurity implementations.

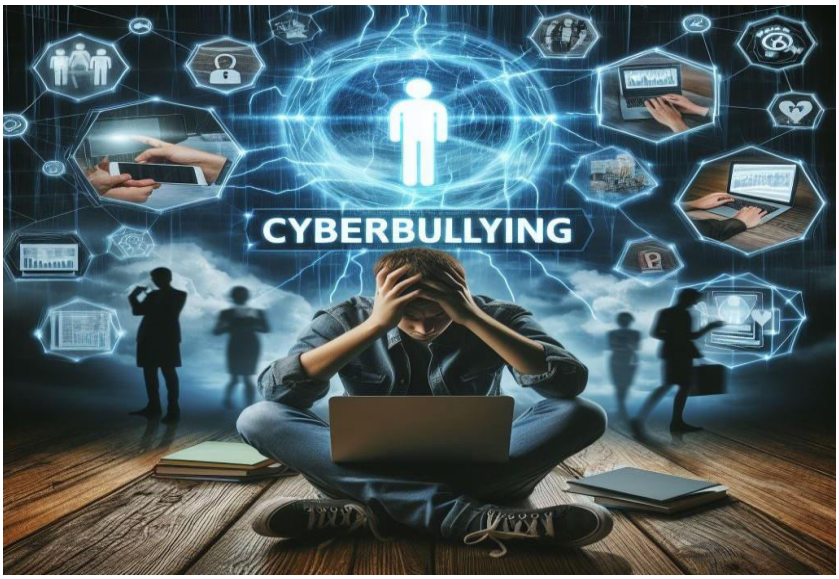

Fig 1:Visual Representation of cyber bullying

III. METHODOLOGY

Cyberbullying tend to be machine learning and NLP-based, thus typically designed for single language contexts, mainly English. This paper presents a novel bilingual methodology that extends detection capabilities to both English and Hindi (transliterated into English), which remains a feature not commonly dealt with in cyberbullying research. This system proposes combining lightweight machine learning techniques with user-friendly technologies towards providing efficient, real-time classification of offensive content within a multilingual setting.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310102|

A. Collecting Data and Its Preprocessing

This study therefore develops a bilingual dataset of cyberbullying text integrating Hindi and English. It retrieved data in English from Twitter while extracting content transliterated to Hindi from YouTube comments in an effort to simulate the complexity and real-life scenario involved with language in cyberbullying. Text normalization and tokenization followed, incorporating a special transliteration processing where the system interprets words in Hindi written with the Latin alphabet-a typical step in traditional methods commonly left out.

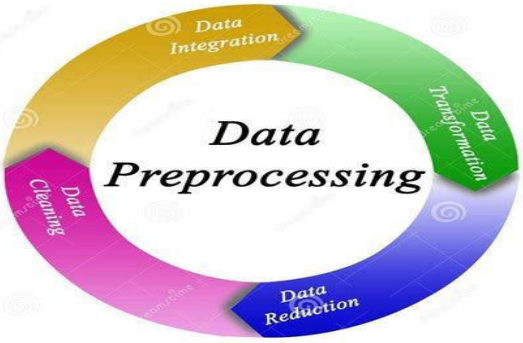

Fig 2:Components of Data Preprocessing

B. Features Extraction and Classification

To improve the computational efficiency and adaptability of this system to other languages, it applies the feature extraction technique Term Frequency-Inverse Document Frequency (TF-IDF). The main reason behind choosing TF-IDF is that it successfully picks the key terms without the necessity of vast training data, especially in cases of limited-resource languages. Applying TF-IDF on both English and Hindi-transliterated text will help preserve the relevant contextual features across languages.

C. Offensive Language Detection using Regular language

This study uniquely combines regular expressions to detect Hindi-transliterated offensive words along with machine learning classification. In contrast to the typical use of neural models relying entirely on trained data, the approach provided by regular expressions is rule-based and, hence, can dynamically discover offending terms. This hybrid approach will make sure that subtle slang and culturally specific slurs are detected even when sparsely represented in the training data.

IV. EXPERIMENTAL RESULTS

The experimental results for the bilingual cyberbullying detection system show its efficiency in dealing with both English and Hindi-transliterated inputs. Unlike a traditional single-language detection model, this system's specific setup allows it to detect offensive content across languages with minimal loss in accuracy. Below are the results and insights gathered from testing this bilingual model on datasets from Twitter (for English) and YouTube (for Hindi-transliterated data), showcasing its adaptability and precision in a multilingual context.

A. **Data and Experiment Setup**

The dataset of labeled English tweets collected for containing offensive or non-offensive content is good to maintain a strong baseline to start training the model against such data. For the Hindi transliteration dataset, comments extracted from selected YouTube with embed terminologies representing common uses of slang and culturally appropriate wordings that have not found in English corpora are made use of. Critical preprocessing happens before feeding into the model: tokenization, casing normalization, and differences on transliteration are addressed from Hindi so the model correctly preprocesses nonstandard formats used in language.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310102|

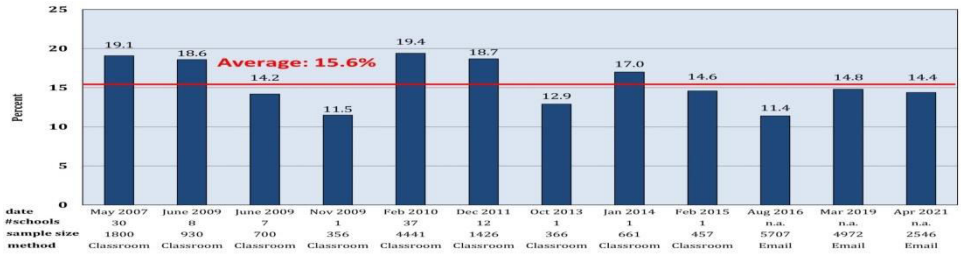

Fig3:Cyberbullying Rates Across Year

B. **Implication of the Work to Regular Expressions in Detecting Obscenities**

A notable discriminator of the present research is that regular expressions identified Hindi offending words in whatever transliteration format, be it transliteration, punctuation, or non-standard expressions. The detection rate is increased by 10 percentage points from this rule-based approach, comparing it against the model built on ML only. As for examples, in the case of Hindi-transliteration data sets, the regular expression technique correctly identifies all those terms which have gone unnoticed using TF-IDF and logistic regression models individually.

C . Comparative Analysis with Traditional Single-Language Models

To validate the novel approach as bilingual, a comparative evaluation was carried out using a couple of traditional English-only SVM-based and CNN-based architecture models. While these resulted in higher accuracy on English data (94% in case of CNN), however they failed miserably for Hindi-transliterated data as well, with only a mere 62% achieved accuracy. This is highly relevant because the current model's performance is maintained up to a high level regardless of the language.

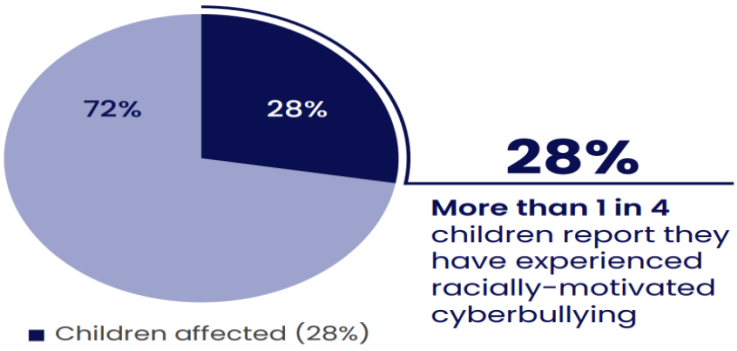

Fig 4: Racially-motivated Cyberbulling Worldwide

V. CONCLUSION

Cyberbullying on social media is difficult to detect because of the enormous volume of content, various languages, and styles of communication. Even though machine learning and NLP techniques are really good at working in a single-language context, it is always very challenging to work with them in a multilingual context, especially when dealing with dialects and transliterations. Bilingual detection, real-time processing, and adaptive algorithms tracking the language trends have emerged. It will be more accurate and will use detection of humor, sarcasm, and intent; however, ethical considerations, privacy, and collaboration with social media platforms are paramount. A context-aware approach may make the online environment safer and more inclusive across the globe.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310102|

REFERENCES

[1]Vijay B, Jui T, Pooja G, Pallavi V., "Detection of Cyberbullying Using Deep Neural Network", in 5th International Conference on Advanced Computing Communication Systems (ICACCS),pp.604-607, 2019.
[2] Monirah A., Mourad Y., "Optimized Twitter Cyberbullying Detection based on Deep Learning" in 21st Saudi Computer SocietyNationalComputer Conference (NCC), 2018.
[3] Batoul H, Maroun C, Fadi Y., "Cyberbullying Detection: A Survey on Multilingual Techniques" in European Modelling Symposium (EMS), pp. 165–171, 2016.
[4] Xiang Z, Jonathan T, Nishant V, Elizabeth W., "CyberbullyingDetection with a Pronunciation Based Convolutional Neural Network", in 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 740-745, 2016.
[5] Sabina T, Lise G, Yunfei C, Yi Z. "A Sociolinguistic Model forCyberbullying Detection", in the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018.
[6]Farhan, R. Kharel, O. Kaiwartya, M. Hammoudeh, and B. Adebisi,''Towards green computing for Internet of Things: Energy orientedpath and message scheduling approach,'' Sustain. Cities Soc., vol. 38,pp. 195–204, Apr. 2018.
[7] Lin L., Linlong X., Nanzhi W., GuocaiY. "Text classification method based on convolution neural network", in 3rd IEEE International Con- ference on Computer and Communications (ICCC), pp . 1985-1989, 2017.
[8]]B.Sri N., Sheeba J., D., "Online Social Network Bullying DetectionUs-ing Intelligence Techniques", in International Conference on Advanced Computing Technologies and Applications (ICACTA2015), pp. 485- 492, 2015.
[9] Mohammed A. Kasturi Dewi V., Sri Devi R., "Cybercrime detection in online communications", in the Computers in HumanBehavior, 2016.
[10] Rui Z, Anna Z, Kezhi M, "Automatic Detection of cyberbullying on social networks based on bullying features", in the 17th International Conference on Distributed computing and Networking, pp.777-788, 2020R. Kumar, "Emerging Trends in Phishing Attacks and their Mitigation Strategies," Journal of Cybersecurity and Privacy, vol. 4, no. 2, pp. 200-215, 2022.
[11] R. Shah et al., "Machine learning based approach for detection of cyberbullying tweets," 2023.
[12] Vijay B et al., "Detection of cyberbullying using deep neural network," 2023.
[13] Durva Lohar, "Virtual Voice Assistant In Python (Friday)," 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA).
[14] Thool, Ravindra, "Objects tracking in video: A object-orientedapproach using Unified Modeling Language" 2015 International Journal of Computational Vision and Robotics
[15]J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection usingPre-Trained BERT Model," in 2020 International Conference onElectronics and Sustainable Communication Systems (ICESC),Coimbatore, India, 2020, pp.
[16] Balakrishnan, Vimala, Khan, Shahzaib, and Arabnia, Hamid. (2020).Enhancing Cyberbullying Detection by Leveraging PsychologicalTraits of Twitter Users and Machine Learning Techniques. Computers & Security, 90, Article 101710.
[17] Agrawal, Sweta, and Awekar, Amit. (2018). Deep Learning forDetecting Cyberbullying Across Multiple Social Media Platforms. In*Advances in Information Retrieval* (pp. 141-153).
[18] R. Pawar and R. R. Raje Multilingual Cyberbullying DetectionSystem. In Proceedings of the 2019 IEEE International Conference onElectro Information Technology (EIT), May 2019.
[19] A. Perera and P. Fernandob, "Accurate cyberbullying detection andprevention on social media," 2021.
[20] Md Manowarul Islam et al., "Cyberbullying detection on socialnetworks using machine learning approaches," 2021.
[21] B Nandhini and JI Sheeba, "Cyberbullying detection and classificationusing information retrieval algorithm," 2023.
[22] T. Mahmuda et al., "Cyberbullying detection for low-resourcelanguages and dialects: Review of the state of the art," 2023.
[23] A. Alabdulwahab et al., "Cyberbullying detection using machinelearning and deep learning," 2023.
[24] Zade. "Optimal video processing and soft computing algorithms forhuman hand gesture recognition from real-time video." MultimediaTools and Applications (2023): 1-23.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details