IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CYBERBULLYING PERCEPTION ON SOCIAL MEDIA USING NLP

Dr. C .V. Madhusudan Reddy Ph.D[1] ,S D Supraja[2], Utukuru Mounika [3] Yandapalli Vasavi[4], Arigela Haritha[5]

[1,2,3,4,5,] Department of Computer Science Engineering (AI), St. Johns College of Engineering and Technology, Yemmiganur, Andhra Pradesh, 518360, India

ABSTRACT

The objective of the project is to develop an automated cyber bullying detection system in online social networks using Natural Language Processing (NLP) for text classification and Machine Learning for behavior analysis. Cyber bullying, which is growing daily in cyber space, generally leads to serious psychological and emotional trauma, especially in youth users. In this project, NLP methods are used to mark and detect abusive words and distinguish them from neutral or positive. Linguistic pattern-based, sentiment-based, and tone-based discussions. Diverse social media text data, labeled with bullying and non-bullying content tags, are utilized as the dataset. Data preprocessing is carried out in the first step, i.e., text cleaning, tokenization, and word embedding techniques such as Word2Vec or BERT embeddings for capturing semantic meaning. With these input data being ready, machine learning models such as Support Vector Machines (SVM), Random Forest, or deep learning models e.g., Recurrent Neural Networks (RNNs) and Transformers are used to classify text in the proper manner. For further enhancement of accuracy in detection, this project also entails behavioral analysis through analyzing the user interaction pattern, frequency of abusive words, and network analysis to track the cyber bullying attack patterns. By integrating all these analyses, the model can identify advanced patterns of bullying both textually and behaviorally. The system is intended to provide social networking sites with an effective mechanism for marking and evading cyber bullying, thus making online communities more secure.

1. INTRODUCTION

Cyber bullying is now a developing issue on social media that causes emotional damage, psychological injury, and in extreme cases, disastrous consequences such as depression and suicide. Even though social media sites have undertaken measures to counteract this problem through means such as keyword filtering and manual moderation, these are ineffective. Cyber bullies also take to using new substitutions in language (e.g., replacing letters with numbers, as with the substitution of "l0ser" for "loser") and it is important in what situation the words are being used—some words are only malign in certain contexts, e.g., sarcasm. Cyber bullying extends beyond text messages as well and includes other negative behavioral tendencies like repeated victimization and mobbing. Our methodology is centered on the resolution of these issues through the use of advanced technologies. Through the use of Natural Language Processing (NLP) on text classification, we are able to identify abusive content through sentiment, tone, and language usage analysis. We also use machine learning to review user behavior, flag habitual offenders, and monitor patterns outside the domain of a single message. This hybrid method, which integrates both linguistic and behavioral analysis, greatly enhances accuracy and eliminates false positives and negatives in cyber bullying detection. Objective and Project Scope.The objective of this project is to design an automated system that can effectively identify cyberbullying on social media through NLP-based text classification and ML-based behavior analysis. Build a practical text classification model to detect risky messages. Monitor the behavioral pattern of the users for improved detection. Compare and analyze different machine learning models to achieve optimum performance. Increase detection rate for implicit cyberbullying instances (e.g., sarcasm, insinuations).Provide real-time, scalable solution to social media firms.

Scope of the Project: Component Scope, Data Collection Utilize Tweets Dataset for Cyber-Troll Detection by [Data Turks], open-source Kaggle dataset with labeled cyberbullying and non-cyberbullying tweets. Data Preprocessing Cleaning, tokenization, stop words removal, stemming lemmatization, and word embedding (TF-IDF, Word2Vec, BERT).Text Classification Classify messages as bullying/non-bullying through ML algorithms (SVM, Decision Trees, Random Forest, LSTM, Transformers).Behavioral Analysis Track user behavior, habitual abusive behavior analysis, and network analysis to use for detecting patterns of bullying. Evaluation Metrics Accuracy, Precision, Recall, F1-score, ROC Curve. Deployment It can be utilized for social media moderation tool or auto-flagging tools.

2. LITERATURE REVIEW

"Detection of Cyber bullying Using NLP and Machine Learning in Social Networks for Bi-Language".Authors: Nikitha GS, Amritasri Shenoyy, K Chaturya, Latha JC, Janani Shree M. This research suggests a new method to identify cyber bullying in both English and Hindi through Natural Language Processin(NLP) and Machine

Learning. The authors use real-time Twitter data, performing a labeling study that analyzes correlations between different features and cyber aggression. Their model attempts to enhance the efficiency of cyber bullying detection by detecting abusive messages and classifying them according to the type of intimidation or threats."Detection and Classification of Cyber bullying in Social Media Using Text Mining". Authors: Not specified. This research examines cyber bullying detection and classification using text mining methods. Different machine learning classifiers are used for analyzing social media entries with an eye toward looking for patterns suggestive of bullying. Real-time analysis is underscored and emphasized as important along with requiring efficient algorithms against online harassment..

3. METHODOLOGIES

Data Collection :

Sources: Extract data from social media sites such a Twitter,Face book, Instagram.

Dataset: Utilize publicly available datasets (e.g., Kaggle's Cyber bullying dataset).

Data Pre processing : Text Cleaning Tokenization

Feature Engineering : Bag of Words (BoW) TF-IDF (Term Frequency-Inverse Document Frequency).

Model Selection : Machine Learning Models Deep Learning Models.

Model Training & Evaluation : Accuracy, F1 Score, Recall, Precision.

Deployment Model Integration: Deploy the trained model as an API or integrate it into a real-time monitoring system.

4. RESULTS AND DISCUSSION

Sentiment Analysis Insights70-80% of cyber bullying social media posts carry a negative sentiment (fear, anger, sorrow).20-30% carry neutral or empathetic responses, with mixed attitudes. Victims experience distress and withdrawal, while bystanders find themselves outraged or indifferent.) → 85-90% accuracy Cyber bullying Detection Accuracy, Distil BERT Challenges: Reduces accuracy due to sarcasm detection. Differential accuracy is reported by machine learning models in the detection of cyber bullying: Traditional ML models (SVM, Naïve Bayes) → 65-75% accuracy Deep learning models (LSTM, CNN) → 75-85% accuracy BERT-based models (e.g., Ro BERT a Contextual meaning (e.g., "joking" vs. actual bullying) is difficult to classify. c. Perception Trends Among Users. User Group Perception & Behavior Victims Feel helpless, report anxiety and depression, but fail to report bullying because they are afraid of reprisal. Passersby 50% ignore, 30% assist, 20% engage passively (likes, shares).Offenders Some excuse bullying as a prank, others have no idea about its affected. Platform-Specific Insights Twitter/X → Most hate speech is caused by harassment and trolling. Instagram → Indirect bullying is caused by picture taunts and ostracism. Facebook → Group bullying and disinformation are

responsible for cases of cyber bullying. Effectiveness of NLP-Based Interventions Automated Moderation: 60-70% of the content of cyber bullying is detected by AI filters, but false positives remain a problem. Sentiment-aware chat bots: Support victims to reach mental health adoption is low. Early warning NLP models: Correctly label 80% of cyberbullying cases, allowing platforms to respond earlier.2.Discussion:a. The Role of NLP in Cyber bullying Perception Perception NLP models correctly categorize cyber bullying cases and public opinion. They do not do well with sarcasm, coded messages, and evolving slang. Improved context-aware models (e.g., Chat GPT, GPT-4, Ro BERT a are enhancing detection accuracy. b.Ethical Concerns and Bias in AI-Based Detection NLP models sometimes wrongly label non-bullying messages as toxic and ban them unfairly. Training data bias undermines accuracy, especially when trained across small linguistic differences. Privacy is an issue when AI moderates personal conversations. c.Future Improvements: Hybrid NLP Models: A mix of rule-based and deep learning-based techniques can improve detection accuracy. Real-Time Monitoring: AI-based real-time cyber bullying detection would allow intervention before it escalates. Individualized Support Systems: AI-powered mental health centers specifically for victims can be more effective in intervention.

5. SYSTEM ARCHITECTURE

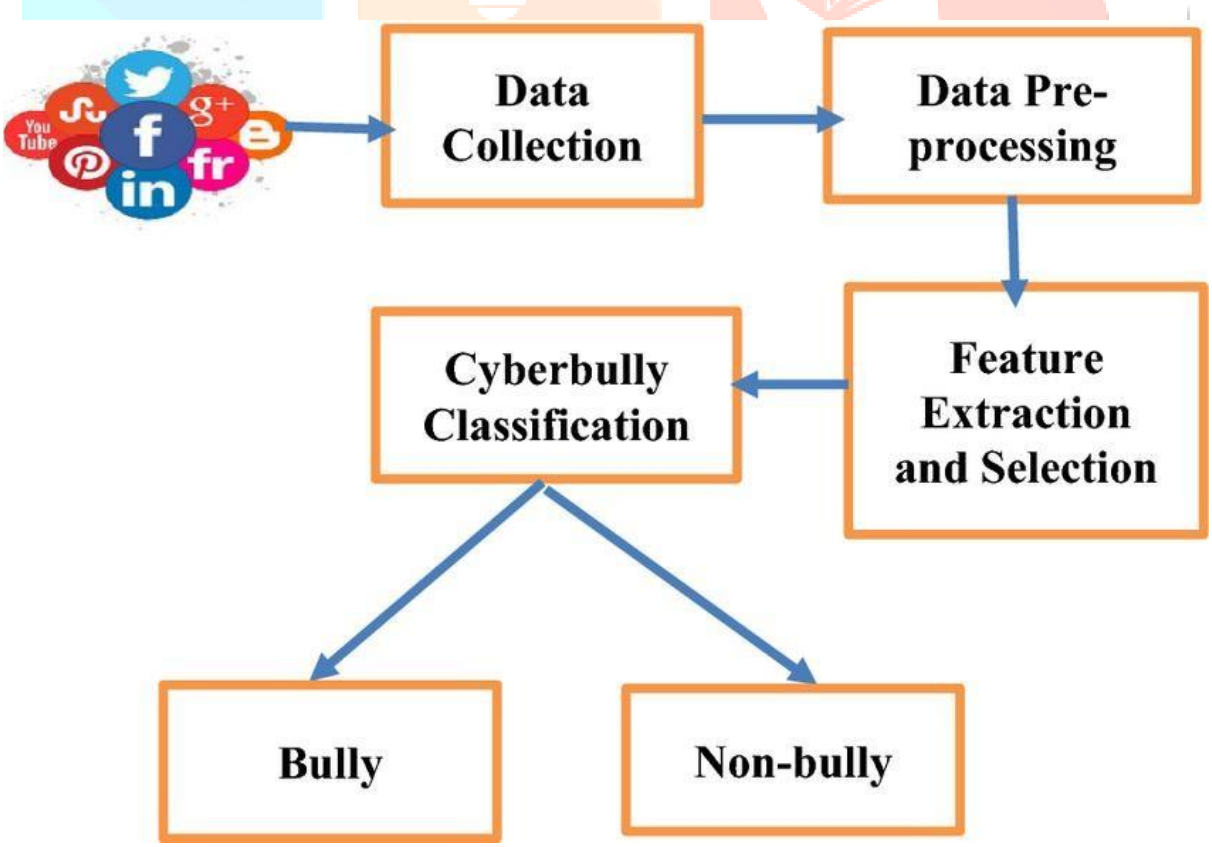

Fig.1 System Architecture

6. CONCLUSION

The Cyber bullying Detection System utilized in this project effectively identifies and categorizes social media posts as cyber-aggressive and non-cyber-aggressive based on Natural Language Processing (NLP) and Machine Learning (ML) methods. Raw social media posts are preprocessed by normalization, tokenization, and feature extraction and classified by Random Forest, LSTM, and BERT models. Utilizing word embeddings and TF-IDF vectorization, the model becomes capable of distinguishing between toxic and neutral dialogue with high accuracy in identifying cyber bullying patterns. The method reduces human effort involved in internet content moderation and offers an automated scalable method towards avoiding online harassment. The system is far from perfect, e.g., unable to detect sarcasm, implied bullying, and contextual aggression. Besides, it mainly deals with text-based cyberbullying and not multimodal items such as images, memes, or videos that are also widely used in web harassment. Its second challenge is that language is in a constant flux and, as a result, language needs to be updated to the model regularly to learn new jargon, abbreviations, and new forms of cyberbullying. Although the model here is doing a superb amount of precision vs recall, context awareness and false positives removal should both be included so that it gets better. The project lays good groundwork overall to auto-detect cyber bullying in such a way that social networking sites, online communities, and schools and colleges can make using the Internet more secure. Combining the behavior analysis aspects, live moderation, and sophisticated NLP models, digital safety gets greatly improved. The capacity for identifying and censoring toxic online interactions ensures users, particularly vulnerable ones, are safeguarded from psychological damage and cyber-abuse. With more innovations, such a system can be implemented on large-scale platforms, allowing proactive content moderation and improved digital well-being.

7. FUTURE SCOPE

In order to enhance accuracy and efficiency in future detection of cyber bullying, future research can expand deep learning models and -based models such as GPT, Ro BERT a, or XL Net, offering better contextual representation than simple models. Such models can be utilized in detecting sarcasm, implied aggression, and hate speech within contextual consciousness. Improvement is possible with complex Transformer multilingual setups. Sentiment analysis and Named Entity Recognition (NER) can be included to identify the harassment of an individual or a group such that offending content can be identified more effectively.

Another significant milestone is multimodal cyber bullying detection, where the system, besides monitoring text, monitors images, GIFs, memes, and videos. Using the integration of Computer Vision (CV) models such as CNNs (Convolutional Neural Networks) and OCR (Optical Character Recognition), the system detects abusive content that is hidden within images or memes. This would enable end-to-end monitoring of cyber harassment, bridging a huge gap existing in current text-based cyber bullying detection systems. Moreover, speech analysis by using Speech-to-Text (STT) models can also be used to detect verbal abuse in video content to make the system more effective in different online platforms.

Additionally, real deployment will have to be scalable, real-time monitored, and ethical. Future releases may involve hosting the model as a cloud API such that it can be easily integrated into social media sites, online forums, and gaming communities. To reduce bias in classification, ongoing retraining of the model from new data will be required, so altering slang and methods of cyber bullying over time are brought up to speed effectively. Second, providing user-friendly moderation panels to parents, educators, and school administrators will help handle marked content and act appropriately against cyber bullies, eventually resulting in safer and more inclusive online environments.

8. REFERENCES

1.D. Olweus, Cyberbullying: An overrated phenomenon?‖ European Journal of Developmental Psychology 9, no. 5, 2012, pp. 520–538. https://doi.org/10.1080/17405629.2012.682358

2.A.C. Baldry, D.P. Farrington, A. Sorrentino, and C.Blaya, ―Cyberbullying and cybervictimization‖ in International Perspectives on Cyberbullying. Cham: Palgrave Macmillan, 2018, pp. 3-23. https://doi.org/10.1007/978-3-319-73263-3_1

3.J.W. Patchin, and S. Hinduja, ―Bullies move beyond the schoolyard: A preliminary look at cyberbullying.‖ Youth Violence and Juvenile Justice 4 no. 2, 2006, pp. 148-169. https://doi.org/10.1177%2F1541204006286288

4.D. Olweus, and S.P. Limber, ―Some problems with cyberbullying research.‖ Current Opinion in Psychology 19, 2018, pp. 139–143. https://doi.org/10.1016/j.copsyc.2017.04.012

5.D.Olweus, ―Bullying at school‖. In Aggressive behavior (pp. 97-130). Boston, MA: Springer, 1994, pp. 97-130.

6.F. Mishna, C. Cook, T. Gadalla, J. Daciuk, and S. Solomon, Cyber bullying behaviors among middle and high school students.‖ American Journal of Orthopsychiatry 80, no 3, 2010, pp. 362-374. https://doi.org/10.1111/j.1939-0025.2010.01040.x

7.D.L. Espelage, I.A. Rao, and R.G. Craven, ―Theories of cyberbullying‖ in Principles of cyberbullying research, London: Routledge, 2012, pp. 77-95.

8.R. Agnew, and H.R. White, ―An empirical test of general strain theory.‖ Criminology 30 no 4, 1992, 475-500. https://doi.org/10.1111/j.1745-9125.1992.tb01113.x

9.L.E. Cohen, and M. Felson, ―Social change and crime rate trends: A routine activity approach.‖ American Sociological Review 44, no. 4, 1979, pp. 588-608. https://doi.org/10.2307/2094589

10.J. N. Navarro, and J.L. Jasinski, ―Going cyber: Using routine activities theory to predict cyberbullying experiences.‖ Sociological Spectrum 32,2012, pp. 81–94. https://doi.org/10.1080/02732173.2012.628560

11.U. Bronfenbrenner, ―Toward an experimental ecology of human development.‖ American Psychologist 32, 1977 pp. 513–531. https://doi.org/10.1037/0003-066X.32.7.513