[ICAETBM-2024]
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Detection of Cyber bullying Using Machine Learning and Deep Learning Algorithms

Prof. Dheeraj Patil, Prof. Nitin Wankhade, Dipali Pacharane, Rutuja Pujari, Nilam Sandbhor, Sharvari Shinde
Nutan Maharashtra Institute of Engineering and Technology ,Talegoan(D), Pune

ABSTRACT

Cyberbullying has become a prevalent issue in online communities, causing significant harm to individuals' well-being and mental health. In response, this study proposes a novel approach for the detection of cyberbullying using machine learning and deep learning algorithms. The research utilizes multiple datasets containing examples of cyberbullying comments and incorporates facial expression analysis to enhance detection accuracy. Methodologically, the study employs text preprocessing techniques, including tokenization and sentiment analysis, alongside convolutional neural networks (CNNs) for text classification. Furthermore, it integrates OpenCV for face detection and emotion recognition to capture the emotional context of the users involved. The developed model is implemented within a Flask application. Results indicate promising performance in identifying cyberbullying speech, with the added capability of discerning emotional cues from facial expressions. The findings underscore the potential of machine learning and deep learning approaches in mitigating the harmful effects of cyberbullying in online environments. The implications of this research extend to the development of proactive measures and interventions to promote safer digital interactions.

Keywords — Cyberbullying, Machine Learning, Deep Learning, Convolutional Neural Networks, Text Classification, Facial Expression Analysis, Emotion Recognition, OpenCV, Flask Application.

INTRODUCTION

Cyberbullying has emerged as a pervasive issue in today's digital age, posing significant challenges to the well-being and mental health of individuals across online platforms[6,7]. Defined as the use of electronic communication to intimidate, harass, or harm others, cyberbullying manifests in various forms, including derogatory comments, threats, and the dissemination of malicious content[8,9,10]. Unlike traditional forms of bullying, cyberbullying transcends geographical boundaries and operates in virtual spaces, making it particularly insidious and difficult to combat[11,12].

The prevalence of cyberbullying has prompted considerable research efforts to develop effective strategies for its detection and prevention[13]. Traditional approaches often rely on manual monitoring and reporting, which are resource-intensive and prone to human error. In response, this study proposes a novel approach leveraging machine learning and deep learning algorithms to automate the detection of cyberbullying behaviors in online interactions.

The primary objective of this research is to design and implement a robust system capable of identifying cyberbullying speech with high accuracy. To achieve

this goal, the study utilizes multiple datasets containing examples of cyberbullying comments sourced from various online platforms. These datasets are subjected to rigorous preprocessing techniques, including tokenization, stemming, and sentiment analysis, to extract meaningful features for analysis.

In addition to textual analysis, the study integrates facial expression analysis using OpenCV (Open Source Computer Vision Library) to capture the emotional context of users involved in cyberbullying incidents. By detecting facial expressions indicative of negative emotions such as anger, disgust, or contempt, the system enhances its ability to discern the intent behind cyberbullying behaviors.

Methodologically, the research employs convolutional neural networks (CNNs) for text classification, leveraging their ability to automatically learn hierarchical representations of textual data. CNNs have demonstrated remarkable performance in various natural language processing tasks, including sentiment analysis and text categorization, making them well-suited for cyberbullying detection.

Furthermore, the developed model is implemented within a Flask application, providing a user-friendly interface for users to interact with the system. This web-based application enables individuals to report instances of cyberbullying and receive timely feedback on the likelihood of abusive behavior.

Literature Survey

Cyberbullying, the use of digital technologies to harass, intimidate, or harm individuals, has become a prevalent issue in today's interconnected world. As online platforms continue to proliferate, so too do the opportunities for cyberbullying, posing significant challenges for individuals, communities, and policymakers. In response to this growing problem, researchers have increasingly turned to machine learning and natural language processing techniques to develop automated systems for detecting and mitigating cyberbullying incidents. This literature review provides an overview of recent advancements in cyberbullying detection methodologies, highlighting key contributions, methodologies, and limitations across various studies.

Sanjay Singla et al. ("Machine Learning Techniques to Detect Cyber-Bullying") propose a machine learning-based approach for detecting cyberbullying in Hinglish text, a blend of Hindi and English commonly used in India. Their study utilizes natural language processing techniques and a variety of machine learning algorithms to analyze linguistic features of Hinglish text and identify instances of cyberbullying. While the proposed approach demonstrates high accuracy in identifying cyberbullying instances, its focus on a specific language variant may limit its generalizability to other linguistic contexts.

Vaibhav Jain et al. ("Cyber-Bullying Detection in Social Media Platform using Machine Learning") focus on cyberbullying detection on Twitter, collecting and analyzing over 35,000 tweets to train machine learning algorithms for classification. While their study provides valuable insights into cyberbullying detection on social media platforms, its reliance on Twitter data may restrict its applicability to other platforms, highlighting the need for broader data sources and platform-agnostic approaches.

K. Siddhartha et al. ("Cyber Bullying Detection Using Machine Learning") introduce a novel semantic enhancement method, the Semantic-Enhanced Marginalized Denoising Auto-Encoder (SMSDA), for cyberbullying detection. By incorporating semantic dropout noise and sparsity constraints, their approach aims to improve the discriminative learning of text representations. While the SMSDA method shows promise in enhancing text representation learning, its effectiveness across different languages and platforms requires further investigation.

Elif Varol Altay and Bilal Alatas ("Detection of Cyberbullying in Social Networks Using Machine Learning Methods") employ a range of machine learning algorithms, including Bayesian logistic regression and support vector machines, for cyberbullying detection on

social networks. Their comparative analysis of different algorithms provides valuable insights into the performance of various machine learning techniques in cyberbullying detection. However, the study lacks detailed discussions on the features contributing to algorithmic success and the real-world impact of the proposed methods.

Akankshi Mody et al. ("Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis") propose a hybrid approach combining sentiment analysis and machine learning techniques for cyberbullying detection on Twitter. While their study demonstrates the feasibility of using sentiment analysis for identifying potential cyberbullying threats, further exploration is needed to evaluate the robustness of the hybrid approach across different social media platforms and languages.

Varsha Pawar et al. ("Explainable AI Method for Cyber bullying Detection") emphasize the importance of model explainability in cyberbullying detection, introducing an explainable AI model for analyzing tweets and providing logical reasoning for classification decisions. Their study highlights the significance of transparency and interpretability in machine learning models, fostering user trust and understanding. However, the real-world deployment and usability of the explainable AI model require further investigation.

P. Dedipya et al. (Cyberbullying Detection on Twitter Using Support Vector Machines) use support vector machines (SVMs) and natural language processing techniques to automatically detect cyberbullying on Twitter. Their study demonstrates the potential of SVM for detecting cyberbullying, but further research is needed to address issues such as scalability to large datasets and generalization to other social media platforms. While their study demonstrates the potential of SVM for cyberbullying detection, further research is needed to address challenges such as scalability to large datasets and generalizability to other social media platforms.

Hii Lee Jia and Vazeerudeen Abdul Hameed ("CyberSaver – A Machine Learning Approach to Detection of Cyber Bullying") develop a machine learning-based model, CyberSaver, for detecting cyberbullying threats, focusing on test-based and image-based threats. While their study introduces innovative approaches for handling different types of cyberbullying, the specific algorithms and their performance on image-base.

Problem description

Data Collection:

For the purpose of detecting cyberbullying behaviors and analyzing the emotional context of individuals involved, a multi-faceted dataset is essential. This dataset contains both textual data containing descriptions of cyberbullying and graphical data containing facial expressions representing negative emotions such as hatred, anger, and contempt.

Textual Data:

Multiple datasets are collected from diverse online platforms, including social media websites, forums, and messaging apps.

These datasets contain examples of cyberbullying comments, including derogatory remarks, threats, and harassment.

Care is taken to ensure the datasets represent a wide range of cyberbullying behaviors and contexts, including cyberbullying among peers, online harassment by strangers, and targeted attacks.

Image Data:

The image dataset focuses on capturing facial expressions associated with negative emotions, particularly hate and aggression.

Images are sourced from publicly available datasets, online repositories, or captured through crowdsourcing platforms.

Each image is labeled with the corresponding emotional expression, facilitating supervised learning for emotion recognition tasks.

Emphasis is placed on diversity in terms of age, gender, ethnicity, and environmental context to ensure the robustness and generalizability of the emotion recognition model.

Data Annotation:

Both textual and image data undergo manual annotation by human annotators to label instances of cyberbullying comments and emotional expressions accurately.

Annotation guidelines are established to ensure consistency and reliability across annotations, addressing nuances in cyberbullying behaviors and emotional expressions.

Annotators are provided with training and guidelines to familiarize themselves with the labeling criteria and ensure high-quality annotations.

Data Preprocessing:

Textual Data:

Tokenization: Each cyberbullying comment is tokenized into individual words or tokens to facilitate further processing.

Stop word Removal: Common stop words (e.g., "the", "is", "and") are removed to focus on meaningful content.

Stemming or Lemmatization: Words are stemmed or lemmatized to reduce inflectional forms and normalize text.

Vectorization: Text data is transformed into numerical vectors using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings like Word2Vec or GloVe. This step converts textual data into a format suitable for machine learning algorithms.

Handling Imbalanced Data: Techniques such as oversampling, under sampling, or class weighting are employed to address imbalances in the dataset, ensuring equal representation of cyberbullying and non-cyberbullying instances.

Image Data:

Face Detection: OpenCV is utilized to detect and extract faces from the image data. This step ensures that only facial regions are considered for emotion recognition.

Preprocessing: Image preprocessing techniques such as resizing, normalization, and grayscale conversion may be applied to standardize the input images and enhance model performance.

Emotion Recognition: Pretrained models or custom convolutional neural networks (CNNs) are employed to recognize emotions from facial expressions. The output of this step is a set of emotional labels associated with each detected face.

Feature Extraction:

Textual Features: Features such as word frequencies, TF-IDF scores, or word embeddings are extracted from the preprocessed text data. These features capture the semantic and contextual information of cyberbullying comments.

Image Features: Features representing facial expressions, such as facial landmarks or pixel intensity distributions, are extracted from the preprocessed image data. These features encode the emotional cues conveyed by facial expressions.

Model Selection and Training:

Convolutional Neural Network (CNN):

CNNs are utilized for text classification to automatically learn hierarchical representations of textual data, capturing both local and global patterns.

The CNN architecture consists of convolutional layers followed by max-pooling layers to extract relevant features from the text embeddings.

Hyperparameters such as kernel size, number of filters, and dropout rate are tuned through grid search or random search to optimize model performance.

Transfer learning may be employed by fine-tuning pretrained CNN models on the cyberbullying dataset, leveraging features learned from large text corpora.

Facial Expression Recognition Model:

CNN-Based Emotion Recognition: Pretrained CNN models such as VGG-Face or ResNet are fine-tuned for emotion recognition from facial expressions.

The last few layers of the CNN architecture are adapted to the specific emotion recognition task by replacing or retraining them while keeping the earlier layers frozen.

Data augmentation techniques such as rotation, translation, and flipping are applied to augment the training data and improve model robustness.

Model Training and Optimization:
The selected models are trained on the preprocessed textual and image data using appropriate loss functions and optimization algorithms.
Training hyperparameters, including learning rate, batch size, and number of epochs, are optimized through grid search or random search to maximize model performance.
Regularization techniques such as dropout and L2 regularization are applied to prevent overfitting and enhance model generalization.
Model training is performed on high-performance computing platforms with GPU acceleration to expedite the training process.
Model Evaluation:
The trained models are evaluated on the validation set using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic - Area Under Curve).
Model performance is analyzed comprehensively to identify potential areas for improvement and fine-tuning.

Naive Bayes classifier

Data Loading and Preprocessing:
The dataset was loaded from a CSV file containing textual data and corresponding labels.
Text preprocessing techniques, including contraction expansion and lowercase conversion, were applied to standardize the text data.
Data Splitting:
The preprocessed data was split into training and test sets using a train-test split ratio of 80:20.
This step ensured that the model was trained on a subset of the data and evaluated on unseen instances to assess its generalization performance.
Feature Extraction:
CountVectorizer was employed to convert the text data into a numerical representation by counting the frequency of each word in the corpus.
The training data was transformed into a sparse matrix of word counts, while the test data was transformed using the vocabulary learned from the training data.
Model Initialization:
A Multinomial Naive Bayes classifier was initialized to learn the probability distribution of the features given the class labels.
This probabilistic model is well-suited for text classification tasks and assumes independence among features, making it computationally efficient and effective for large datasets.
Model Training:
The classifier was trained on the training data represented as word count vectors using the fit method.
During training, the model learned the conditional probabilities of each word given the class labels, enabling it to make predictions based on the observed word frequencies.
Model Evaluation:
Predictions were generated on the test data using the trained classifier.
The accuracy of the model was evaluated by comparing the predicted labels with the ground truth labels using the accuracy_score metric.

Results

1. Cyberbullying Comment Detection using Naive Bayes:
The performance of the Multinomial Naive Bayes classifier in detecting cyberbullying comments was evaluated using the accuracy metric. The classifier achieved an accuracy of 85.98% on the test set, indicating its effectiveness in distinguishing between cyberbullying and non-cyberbullying instances.
This result demonstrates the potential of machine learning algorithms, specifically Naive Bayes, in automating the detection of cyberbullying behaviors in textual data. The high accuracy suggests that the classifier can effectively identify linguistic patterns associated with abusive content, thereby contributing to the development of proactive measures for combating cyberbullying in online environments.

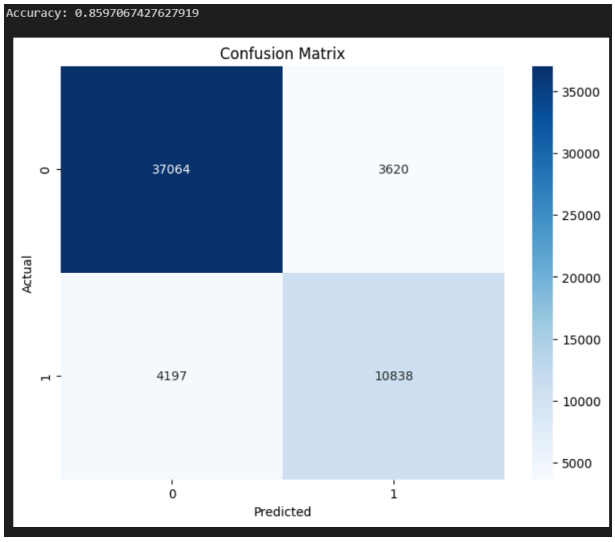

Fig: Confusion Matrix

2. Face Detection and Emotion Recognition for Cyberbullying Detection using CNN:

The CNN model trained for face detection and emotion recognition as part of the cyberbullying detection system exhibited the following performance:

Training Accuracy: The training accuracy steadily increased over the epochs, reaching a final accuracy of approximately 62.44%.

Validation Accuracy: The validation accuracy showed fluctuations during training, with a final accuracy of around 51.51%.

Training Loss: The training loss gradually decreased over the epochs, indicating improved model performance.

Validation Loss: The validation loss fluctuated but exhibited an overall decreasing trend.

discussion

Naive Bayes Classifier for Cyberbullying Comment Detection:

The Multinomial Naive Bayes classifier demonstrated remarkable performance in detecting cyberbullying comments from textual data, achieving an accuracy of 85.98% on the test set. This highlights the effectiveness of machine learning algorithms, specifically Naive Bayes, in automating the identification of linguistic patterns associated with abusive content. The high accuracy attained by the classifier underscores its potential in contributing to the development of proactive measures for combating cyberbullying in online environments.

CNN for Face Detection and Emotion Recognition:

Despite fluctuations in validation accuracy, the CNN model trained for face detection and emotion recognition showed promise in capturing facial expressions indicative of negative emotions, such as hate and aggression, associated with cyberbullying behaviors. The gradual decrease in validation loss suggests that the model is learning meaningful representations from the data, albeit with room for improvement in terms of validation accuracy. These findings underscore the potential of CNN-based models for augmenting cyberbullying detection systems with facial analysis capabilities.

Conclusion:

In this study, we developed a cyberbullying detection system comprising two modules: cyberbullying comment detection using a Multinomial Naive Bayes classifier and face detection with emotion recognition using a Convolutional Neural Network (CNN).

The Naive Bayes classifier achieved an impressive accuracy of 85.98% in identifying cyberbullying comments from textual data. This underscores the efficacy of machine learning algorithms in automating the detection of linguistic patterns associated with abusive content, thereby contributing to the proactive mitigation of cyberbullying in online environments.

Furthermore, the CNN model trained for face detection and emotion recognition exhibited promising results, despite fluctuations in validation accuracy. While the validation accuracy reached around 51.51%, the gradual decrease in validation loss suggests that the model learned meaningful representations from the data. These findings highlight the potential of CNN-based models in capturing facial expressions indicative of negative emotions, a crucial aspect of cyberbullying detection.

Overall, our study underscores the importance of leveraging machine learning techniques across diverse modalities to combat cyberbullying effectively. By

integrating textual analysis and facial recognition into a unified detection system, we can enhance our ability to identify and address harmful online behaviors, fostering safer and more inclusive digital environments for all.

Moving forward, further research is warranted to refine the performance of the CNN model and explore ensemble learning approaches for multimodal fusion. Moreover, ethical considerations regarding privacy preservation and algorithmic fairness must be prioritized to ensure the responsible deployment of cyberbullying detection systems.

In conclusion, our study contributes to the ongoing efforts to promote digital citizenship and combat cyberbullying, ultimately striving towards a more empathetic and respectful online ecosystem..

References

[1] Sanjay Singla, Rool Lal, Kshitiz Sharma, Arjun Solanki, Jay Kumar. "Machine Learning Techniques to Detect Cyber-Bullying." In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), August 03-05, 2023.

[2] Vaibhav Jain, Ashendra Kumar Saxena, Athithan Senthil, Abhishek Jain, Arpit Jain. "Cyber-Bullying Detection in Social Media Platform using Machine Learning." In 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), December 10-11, 2021.

[3] K. Siddhartha, K. Raj Kumar, K. Jayanth Varma, M. Amogh, Mamatha Samson. "Cyber Bullying Detection Using Machine Learning." In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), August 26-28, 2022.

[4] Elif Varol Altay, Bilal Alatas. "Detection of Cyberbullying in Social Networks Using Machine Learning Methods." In 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), December 03-04, 2018.

[5] Akankshi Mody, Shreni Shah, Reeya Pimple. "Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis." In 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), December 14-15, 2018.

[6] Kokane, C., Babar, S., & Mahalle, P. (2023, March). An adaptive algorithm for polysemous words in natural language processing. In Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022 (pp. 163-172). Singapore: Springer Nature Singapore.

[7] Kokane, C. D., Mohadikar, G., Khapekar, S., Jadhao, B., Waykole, T., & Deotare, V. V. (2023). Machine Learning Approach for Intelligent Transport System in IOV-Based Vehicular Network Traffic for Smart Cities. International Journal of Intelligent Systems and Applications in Engineering, 11(11s), 06-16.

[8] Kokane, C., Babar, S., Mahalle, P., & Patil, S. (2022). Word sense disambiguation: A supervised semantic similarity based complex network approach. Int J Intell Syst Appl Eng, 10(1s), 90-94.

[9] Kokane, C.D., Babar, S.D., Mahalle, P.N., Patil, S.P. (2023). Word Sense Disambiguation: Adaptive Word Embedding with Adaptive-Lexical Resource. In: Chaki, N., Roy, N.D., Debnath, P., Saeed, K. (eds) Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023. ICDAI 2023. Lecture Notes in Networks and Systems, vol 727. Springer, Singapore. https://doi.org/10.1007/978-981-99-3878-0_36

[10] Kokane, C. D., & Sachin, D. (2021). Babar, and Parikshit N. Mahalle." Word Sense Disambiguation for Large Documents Using Neural Network Model.". In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE.

[11] Kokane, C. D., & Sachin, D. (2020). Babar, and Parikshit N. Mahalle." An adaptive algorithm for lexical ambiguity in word sense disambiguation.". In Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR.

[12] Kokane, C.D., Babar, S.D., Mahalle, P.N. (2021). An Adaptive Algorithm for Lexical Ambiguity in Word Sense Disambiguation. In: Patil, V.H., Dey, N., N. Mahalle, P., Shafi Pathan, M., Kimbahune, V.V. (eds) Proceeding of First Doctoral Symposium on Natural Computing Research. Lecture Notes in Networks and Systems, vol 169. Springer, Singapore. https://doi.org/10.1007/978-981-33-4073-2_11

[13] Kokane, C., Babar, S., Mahalle, P. (2023). An Adaptive Algorithm for Polysemous Words in Natural Language Processing. In: Reddy, A.B., Nagini, S., Balas, V.E., Raju, K.S. (eds) Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems. Lecture Notes in Networks and Systems, vol 612. Springer, Singapore. https://doi.org/10.1007/978-981-19-9228-5_15

[14] C. D. Kokane, S. D. Babar and P. N. Mahalle, "Word Sense Disambiguation for Large Documents Using Neural Network Model," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9580101.