Secure Social Media with Image and Text Moderation

Dr.V.Kavitha[1], Ms.R.Kavyasri[2], Ms.K.Kaviyadharsini[3], Mr.S.Krishna Kumar[4]
[1]AssociateProfessor, [2,3,4]Student, Department Of Computer Science And Engineering
Velalar College Of Engineering And Technology, Erode

ABSTRACT - Social media platforms such as Facebook, Twitter, and WhatsApp play a pivotal role in modern communication, facilitating the sharing of personal interests, activities, and daily conversations. With the growing significance of these platforms, ensuring the privacy and security of user-generated content has become a paramount concern. This paper addresses the challenges associated with safeguarding sensitive information on social media applications and proposes an advanced data protection method. The primary objective of this research is to develop a robust framework that combines Convolutional Neural Network (CNN) algorithms, supervised machine learning (ML) techniques, and the TensorFlow ML framework to enhance the privacy and security of user-shared content. The proposed framework leverages advanced permission settings, enabling users to control access levels for specific files and information, thereby ensuring that only authorized individuals or groups can view or modify sensitive content. The research focuses on the integration of CNN algorithms to analyze and classify user-generated content, employing supervised learning techniques to continuously refine the system's ability to identify and protect sensitive information. The TensorFlow ML framework is utilized to enhance the efficiency and scalability of the proposed data protection method, ensuring seamless integration into existing social media architectures.

1. INTRODUCTION

In addition to personal use, these applications also serve various purposes such as business, education, and trading chats. However, despite their positive aspects, there are also negative aspects to consider. The existence of vulgarities on these platforms is a cause for concern. Students, in particular, are vulnerable to these vulgarities, and exposure to such content can lead them astray, negatively impacting their future prospects. Pornographic or adult-oriented materials, for example, are often shared without regard for their harmful effects. Regular dull in comparison to the excitement generated by pornography, making it even more challenging to concentrate on them. To prevent such outcomes, social media applications should prohibit the sharing of such content. In today's technology-driven era, traditional communication methods such as SMS or phone calls are being used less frequently and are being replaced by chat applications. Social media platforms have become an integral part of everyday life. Among the most popular chat applications today is WhatsApp. With WhatsApp, users have various options for communicating, including sending text messages and . making phone calls. Additionally, the app allows for sharing of photos and videos. The rise of social network platforms has led to unprecedented levels of connectivity among individuals. As these platforms continue to evolve at an accelerated pace, even into the era of virtual reality, our privacy becomes increasingly vulnerable to a growing number of threats. One of these threats is the proliferation of vulgar and inappropriate content, particularly in the form of shared images, which can have harmful or sexual content. This Application is developed with the factors that ensures the user's privacy and helps in making secure conversations. Image moderation is a type of content moderation that aims to screen out images that are deemed explicit or unsuitable for a brand's social media platform. The number of images that are assessed by the image moderation system is considered as the measure of image moderation. The image moderation process functions by identifying the links within messages and transmitting the content of those links to our AI moderation service for analysis. TensorFlow is a comprehensive open-source platform for machine learning that handles all facets of a machine learning system. However, in this course, emphasis is placed on using a specific TensorFlow API to build and train machine learning models. TensorFlow facilitates the creation of dataflow graphs and structures that specify how data is processed via a multi-dimensional array called a Tensor. This permits you to design a series of operations that can be performed on the input and output values, thereby constructing a flowchart.

Developing a Convolutional Neural Network (CNN) for image classification to avoid harmful or inappropriate content is a crucial aspect of many projects, especially those involving user-generated content. Below is an overview of the key components and considerations for implementing a CNN for this purpose:

1.1 DATASET

Positive Samples: Collect a dataset of images containing harmful or inappropriate content. This serves as the positive class.

Negative Samples: Gather a diverse set of images that represent normal, safe content. This will be the negative class.

Ensure a balanced and representative dataset for effective model training.

1.2 DATA PREPROCESSING

Resize images to a consistent size for uniformity. Normalize pixel values to a specific range (e.g., [0, 1]). Augment data through techniques like rotation, flipping, and zooming to increase model robustness.

1.3 MODEL ARCHITECTURE

Design a CNN architecture suitable for image classification. Common architectures include VGG, ResNet, or custom-designed models. Consider using pre-trained models to leverage learned features from large datasets (e.g., ImageNet). Add layers like Convolutional, Pooling, Batch Normalization, and Dense layers.

1.4 TRAINING

Split the dataset into training and validation sets. Train the model on the training set and validate it on the validation set. Monitor metrics like accuracy, precision, recall, and F1 score.

The effectiveness of the CNN depends on the quality of the dataset and the careful tuning of hyperparameters. Regular updates and continuous monitoring are essential for maintaining the model's accuracy over time.

2. LITERATURE REVIEW

2.1 Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media

ML algorithms for hate speech detection in social media, addressing key components like data collection, feature extraction, and model evaluation. It contributes by informing readers on detection steps, evaluating method strengths and weaknesses, and identifying research gaps, providing valuable insights for researchers and professionals in the field.

2.2 Automatic Detection of Offensive Language for Urdu and Roman Urdu

Challenge of detecting offensive language on social media in the Urdu language, introducing the first offensive dataset for Urdu comments. Using character-level n-grams and various machine learning techniques, the research achieves superior performance, with LogitBoost and SimpleLogistic reaching 99.2% and 95.9% F-measure on Roman Urdu and Urdu datasets, respectively.

2.3 Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data

Introduces annotation guidelines, a new dataset (RU-HSD-30K), and a context-aware a hate speech detection model employing a Bi-LSTM architecture with an incorporated attention layer in Roman Urdu, outperforming traditional and outperforming alternative deep learning models with an accuracy reaching 0.875 and an F-Score of 0.885, while showcasing enhanced performance through lexical normalization.

2.4 A Framework for Hate Speech Detection Using Deep Convolutional Neural Network

Addresses the challenge of hate speech on Twitter, proposing an automated system based on Deep Convolutional Neural Network (DCNN) with GloVe embedding vectors. The model achieves high precision (0.97), recall (0.88), and F1-score (0.92), outperforming existing models and offering efficient hate speech detection in the vast volume of tweets.

2.5 Political Hate Speech Detection and Lexicon Building: A Study in Taiwan

Proposes a comprehensive approach for detecting hate speech in Chinese, involving the creation of developing an artificial intelligence classifier by training it on a political hate speech lexicon, and developing both deep learning and lexicon-based methods. The system and lexicon methodology are adaptable for detecting hate speech in various languages, addressing the challenges of maintaining online freedom of speech while combatting hate speech proliferation.

2.6 G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media

Proposing G-BERT, a novel model combining BERT and GRU, achieves high accuracy (95.56%) in detecting hate speech in Bengali social media posts, outperforming other algorithms and demonstrating effectiveness in mitigating online hate speech.

2.7 Pornography object detection using Viola-Jones algorithm and skin detection

Introducing an application employing the Viola-Jones algorithm along with skin detection to detect pornography objects, achieving a test accuracy of 78.29% with optimized parameters, addressing the need for preventing exposure, especially for children.

3. EXISTING SYSTEM

In today's technology-driven era, traditional communication methods such as SMS or phone calls are being used less frequently and are being replaced by chat applications. Social media platforms have become an integral part of everyday life. Among the most popular chat applications today is WhatsApp. With WhatsApp, users have various options for communicating, including sending text messages and making phone calls. Additionally, the app allows for sharing of photos and videos. The rise of social network platforms has led to unprecedented levels of connectivity among individuals. As these platforms continue to evolve at an accelerated pace, even into the era of virtual reality, our privacy becomes increasingly vulnerable to a growing number of threats. One of these threats is the proliferation of vulgar and inappropriate content, particularly in the form of shared images, which can have harmful or sexual content.

There is no option available to block specific users from viewing your profile. Messages are not delivered to the regular mobile phone inbox. There is a risk of unauthorized access to personal messages, causing problems in personal relationships. Previously, the limit for group members was 256, but it can now be increased to 3000 by changing a few settings. Frequent message notifications can be irritating to some individuals. Yourprofile picture is visible to anyone who has your contact number saved on WhatsApp. WhatsApp has addictive qualities, particularly among school children, and breaking free from addiction can be challenging. You must share your phone number with those you wish to communicate with on WhatsApp.

4. PROPOSED SYSTEM

This Application is developed with the factors that ensures the user's privacy and helps in making secure conversations.

4.1 Image Moderation:

Image moderation is a type of content moderation that aims to screen out images that are deemed explicit or unsuitable for a brand's social media platform. The number of images that are assessed by the image moderation system is considered as the measure of image moderation. The image moderation process functions by identifying the links within messages and transmitting the content of those links to our AI moderation service for analysis.

4.2 TensorFlow:

TensorFlow is a comprehensive open-source platform for machine learning that handles all facets of a machine learning system. However, in this course, emphasis is placed on using a specific TensorFlow API to build and train machine learning models. TensorFlow facilitates the creation of dataflow graphs and structures that specify how data is processed via a multi-dimensional array called a Tensor. This permits you to design a series of operations that can be performed on the input and output values, thereby constructing a flowchart.

4.3 TensorFlow Architecture:

Tensorflow architecture works in three parts:

- Pre-processing the data
- Build the mode
- Train and estimate the model

The term "TensorFlow" is derived from the fact that it receives input in the form of a multidimensional array, commonly referred to as tensors. By defining a series of operations that you want to perform on this input, you can create a flowchart, known as a Graph. The input is inserted at one end, flows through the system of multiple operations, and emerges at the other end as output. TensorFlow has gained significant recognition as the leading deep learning library in recent years. Users of TensorFlow can create various deep learning architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and even simple artificial neural networks. Academics, startups, and major corporations are among the primary users of TensorFlow. Google incorporates TensorFlow in nearly all of its daily products, including Gmail, Google Photos, and the Google Search Engine.

Advantages of the proposed systems are:

- Privacy among users will be ensured
- Misleading will be avoided
- Useful and harmless to use for all age people
- Nomalicious contents allowed

The working architecture of the chat application is clearly represented using diagrammatic format. For the front end of the application flutter is used. After authentication the user is allowed to chat. The chat can be personal and be a group chat. The user can also share their media files. The media files undergo image moderation and after that will be sent. All the chats will be stored in the chat history.

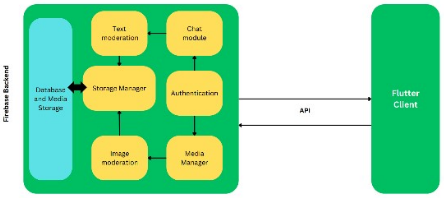

Figure 1 : System Architecture

The ratings of sexual content showed a significant main effect of gender, with females rating compared to males. Table provides a summary of the mean ratings by males and females.

	Existing System		Proposed System	
Rating of:	Males Mean (SD)	Females Mean (SD)	Males Mean (SD)	Females Mean (SD)
Sexual content—funny	3.53 (1.80)	2.04 (2.02)	0	0
Sexual content—exciting	2.75 (1.85)	1.08 (1.43)	0	0
Sexual content—disturbing	3.32 (1.49)	4.56 (1.66)	0	0
Violent content—funny	1.55 (1.66)	0.71 (1.35)	0	0
Violent content—exciting	1.42 (1.62)	0.60 (1.13)	0	0
Violent content—disturbing	4.21 (1.83)	4.99 (1.72)	0	0

This application presents a table featuring a rating of zero, indicating that there will be no impact on the sexual content generated.

As shown in Figure 2, over 2/3 of respondents had received explicit sexual content, with no major differences between males and females in each category. Fewer respondents overall had received explicit violent content(in Figure 3), and distribution of females across categories here was significantly different from males; over 60% of males had received violent content, compared with just 40% of females.

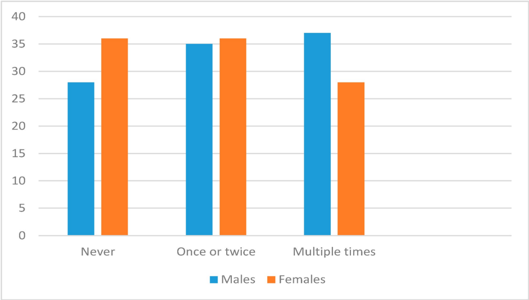

Figure 2 : Frequency of prior exposure to unsolicited sexual content by males and females

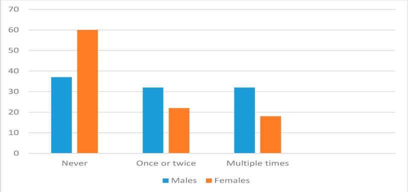

Figure 3 : Frequency of prior exposure to unsolicited violent content by males and females

After the source code has been completed, documented as related data structures. Completed the project has to undergo testing and validation where there is subtitle and definite attempt to get errors.

The project developer tread lightly, designing an execution test that will demonstrate that the program works rather than uncovering errors, unfortunately errors will be present and if the project developer doesn't find them, the user will find out.

The project developer is always responsible for testing the individual units i.e. modules of the program. In many cases the developer also conducts integration testing i.e. the testing step that leads to the construction of the complete program structure.

This project has undergone the following testing procedures to ensure its correctness.

- Unit Testing
- User Acceptance Testing

5. CONCLUSION AND FUTURE ENHANCEMENT

In this work, proposed the Chat application which provides privacy and security among users through prohibiting sharing the harmful contents that are sexual contents. This project will improve the usage of applications with the help of secure content sharing with safety. In this, the sexual content sharing is prohibited with the help of above mentioned tools. This chat application will also ensure the safety and security of students who are using this application.

In conclusion, a platform for sharing and collaboration on sensitive content with advanced permissions and access control for users is crucial for organizations and individuals that deal with sensitive information. The platform can provide a secure environment for users to collaborate and share sensitive content, while also ensuring that access is limited to authorized individuals. With advanced permission settings, users can control who can access, modify, or delete content, reducing the risk of data breaches or unauthorized access. This platform can improve the efficiency and security of sensitive data management, while also enhancing collaboration and communication between teams. By implementing such a platform, organizations and individuals can protect their sensitive information and maintain confidentiality, which is critical for maintaining trust and credibility in today's digital age.

REFERENCES

[1] N. s. m. and W. m. n. w. z. , "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," vol. 9, pp. 88364-88376, 2021.

[2] M. p. a. Z. j. I. r. n. M. a. and M. t. s. , "Automatic Detection of Offensive Language for Urdu and Roman Urdu," vol. 8, pp. 91213-91226, 2020.

[3] F. r.-s. J. c.-d.-a. and L. p. , "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data," vol. 8, pp. 219563-219576, 2020.

[4] C.Nagarajan and M.Madheswaran - ‘Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter’ - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.

[5] C.Nagarajan and M.Madheswaran - ‘Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis’- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011

[6] C.Nagarajan and M.Madheswaran - ‘Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques’- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.

[7] C.Nagarajan and M.Madheswaran - ‘Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis’- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.

[8] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. “Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter” Journal of VLSI Design Tools &amp; Technology. 2022; 12(2): 34–41p.

[9] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya “Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay” Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.

[10] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash “Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM” Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756

[11] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, “Perfromance Investigation of T-Source Inverter fed with Solar Cell” Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749

[12] C.Nagarajan and M.Madheswaran, “Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation” has been presented in ICTES’08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007

[13] M Suganthi, N Ramesh, “Treatment of water using natural zeolite as membrane filter”, Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022

[14] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, “Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District”, International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019