DETECTING OF ILLICIT MESSAGES AND IMAGES ON SOCIAL MEDIA

PALAGANI JAGATHI, PENTAPALLI MEGHANA, SIRIPURAPU SURESH, YAMPARALA SATISH, Student, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

Ms. T. SYAMA SREE, Assistant Professor, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P, India.

ABSTRACT

Online Social Networks is to allow people to communicate virtually by using the internet. The rise in web and social media interactions has resulted in the effortless proliferation of offensive language and hate speech. This action involves repeated online insulting, harassing, or attacking a target verbally.These content adversely affecting their mental health, and demeaning the integrity of social networking platforms. Illicit content such as hate speech, offensive content is a growing concern on social media platforms. To address this issue, various techniques have been developed. These methods typically involve the use of machine learning algorithms trained on large datasets of labeled examples to identify patterns inthe text and images that indicate illicit content. Some popular techniques include natural language processing (NLP) for identifying hate speech and other offensive language, and deep learning models for identifying patterns in images that may be legal or illegal. NLP techniques are applied to analyze text- based social media posts to identify patterns and features that may indicate illicit content. An XGBoost classifier is trained using these features to make predictions. CNN are used to process and analyze images, where a trained model is applied to identify illicit images. The system is evaluated using a dataset of social media posts and images, showing a high accuracy in detecting illicit content.

INTRODUCTION

Illicit content such as hate speech, and child exploitation, human trafficking is a growing concern on social media platforms. Detecting and removing this type of content is essential to maintain a safe and positive environment for users. One approach to detecting illicit content is to use machine learning algorithms, such as XG Boost and CNN. XG Boost is a powerful algorithm that can be used to classify text-based messages, while CNN is a deep learning algorithm that excels at image classification tasks. once the model is trained, it can be used to detect and flag potentially problematic content on social media platforms, helping to keep users safe and promote a positive online community. It is important to note that these models may have limitations in terms of accuracy and bias, and also that the detection of illicit messages and images is a complex task.

Illicit content detection on social media involves using natural language processing (NLP) and machine learning techniques to identify and flag potentially harmful or illegal content, such as hate speech, cyberbullying, or child exploitation. One popular approach is to use a combination of NLP techniques and a machine learning algorithm, such as XGBoost or a convolutional neural network (CNN), to analyze text and images and detect patterns that indicate illicit content. XGBoost is a gradient boosting algorithm that is often used for classification and regression tasks. It is known for its efficiency and ability to handle large amounts of data. In the context of illicit content detection, XGBoost can be used to analyze text and identify patterns that indicate hate speech or other forms of harmful content.

CNNs are a type of neural network that are commonly used for image classification tasks. They can be used to detect patterns in images and identify objects or features that may indicate illicit content. In the context of illicit content detection, CNNs can be used to analyze images and detect patterns that indicate child exploitation or otherforms of harmful content.

LITERATURE SURVEY

- One study published in 2018 used an XGBoost classifier to detect hate speech on Twitter. The classifier was trained on a dataset of tweets labeled as hate speech or not, and achieved an accuracy of 85.3%.
- Another study published in 2019 used a combination of NLP and CNNs to detect abused explicit content on Instagram. The model first used NLP to extract text features from captions and comments, and then used a CNN to analyze the images.The model achieved an F1-score of 0.94 in detecting abused explicit content.
- Detecting Illicit Content on Social Media: A Survey of Text and Image Analysis

Techniques" by A. F. S. T. Abrar, M. Imran, and M. A. Imran. Published in IEEE Communications Surveys & Tutorials, 2019. This paper provides an overview of the state-of-the-art techniques for detecting illicit content on social media, with a focus on text and image analysis.

- "Deep Learning for Illicit Content Detection on Social Media" by M. A. Imran, A.

  F. S. T. Abrar, and M. Imran. Published in IEEE Access, 2018. This paper presents a survey of recent deep learning-based methods for detecting illicit content on social media, including text and image analysis.

- "Machine Learning for Detecting Illicit Content on Social Media: A Survey" by A.

  F. S. T. Abrar, M. Imran, and M. A. Imran. Published in IEEE Communications Surveys & Tutorials, 2017.

- Granizo, Sergio & Alvarez, Myriam & Barona, Lorena & Valdivieso, Leonardo. (2020). Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. IEEE Access. PP. 1-1.

  In this paper, we identify Twitter messages that could promote these illegal services and exploit minors by using natural language processing. The images and the URLs found in suspicious messages were processed and classified by gender and age

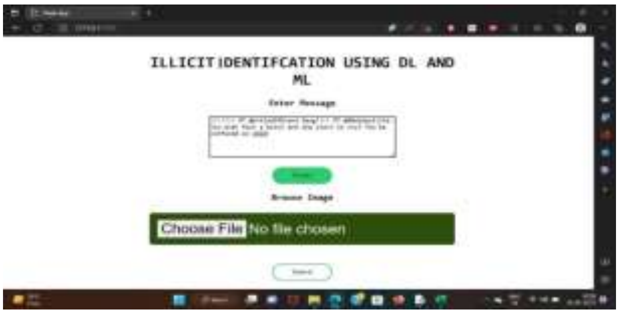
  group, so it is possible to detect photographs of people.

EXISTING SYSTEM

In recent years the illegal activities became more in the social media. So the negativity increases on these social media platforms. In existing system , the project can able to detect the illicit messages that are posted on social media like face book ,twitter , Instagram , etc.., Machine learning and deep learning is used to find the illegal messages.These algorithms were used to find those illegal messages and the user who posted them on social media also be recognized . Support Vector Machine (SVM) is used to identify the illegal messages and Convolution neural network (CNN) is used to identify the illegal images on the social media platforms. These illegal messages and images are identified with some accuracy . Gender and age is identified by extracting the features of the image and with the help of convolution neural network. we can identify them with some accuracy. This project can train only smaller datasets. So, We get the low accuracy with the smaller dataset.

PROPOSED SYSTEM

As fast as technology grows crime rate also grows in different ways. To decrease the crime rate we are detecting illegal messages and images that are posted on social media. The system would first need to collect a large dataset of text and images from social media platforms that can be used to train the machine learning models. This dataset would need to include a mix of both legitimate and illicit content. The database contains the trained data which is used to identify whether the message is hate speech, neither or offensive language. Not only messages, images are also classified as illegal, neutral and abused. The messages can be preprocessed and then identified them as their respective categories. The images can be extracted with its features using convolution neural network (CNN) and classified them as those categories. Then the machine learning and deep learning algorithms are used to find the illegal messages and images. The accuracy of the illegal words and images are also identified. The F1 score is showed between the range 0 to 1 and accuracy of the project is 88%.We have used XG boosting algorithm to get the best accuracy. So, this project can helps to stop posting the unauthorized messages and images on social media platforms. When we are predicting a sentence it is tokenized and spilts into single word and give a mean value and with a score.

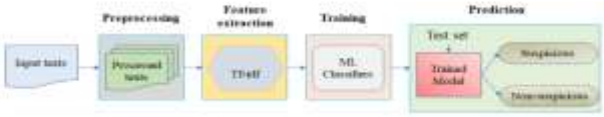

ILLICIT IDENTIFCATION USING DL AND ML

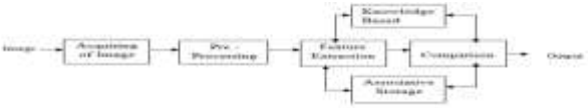
Enter Message

Browse Image

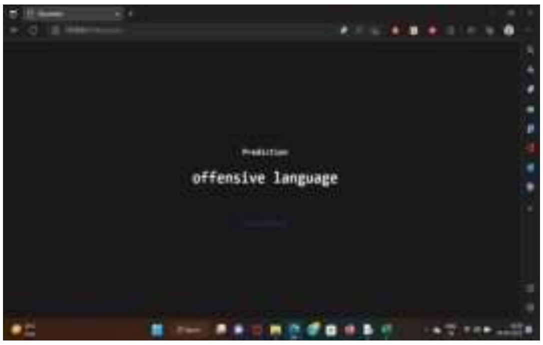
Choose File No file chosen

Process Diagram



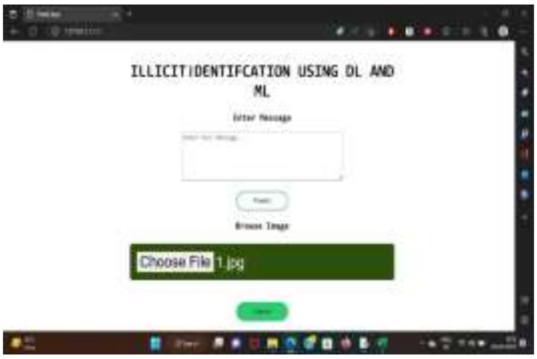


SAMPLE RESULTS



Prediction

offensive language

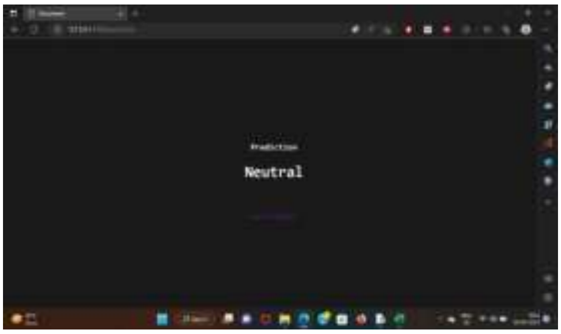


ILLICITIDENTIFCATION USING DL AND ML

Enter Message

Browse Image

Choose File 1.jpg



Prediction

Neutral

CONCLUSION

In conclusion, detecting illicit messages and images on social media is a challenging task that requires a combination of advanced technologies and human oversight. Natural language processing (NLP) and machine learning algorithms such as XG Boost and convolutional neural networks (CNNs) can be used to analyze text and images and identify patterns and features that indicate inappropriate content. Therefore, it is important to have a robust dataset to train the model on and to have a human oversight to check the results. Additionally, it is important to have a clear and consistent policy in place for handling and removing inappropriate content from social media platforms. In all, The detection of illicit messages and images is an ongoing process that requires a combination of technology and human effort to ensure asafe and healthy online environment.

FUTURE ENHANCEMENT

As the technology changes or new requirements are expected by the user, to enhance the functionality of the product may require new versions to be introduced. But we mainly made this project to reduce the negativity on social media.We come up with a new application which detect both illegal messages and images on social media platforms.We can also predict the age through the use other algorithms in future,Although the system is complete and working efficiently, new changes which enhance the system functionality can be added without any major changes to the entire system.

REFERENCES

1. [1] Rybnicek M, Poisel R, Tjoa S,(2013) Facebook Watchdog: Research Agenda for Detecting Online Grooming and Bullying Activities, Systems, Man, and Cybernetics (SMC),2013.

[2] Reynolds, Kelly, April Kontostathis, and Lynne Edwards.(2011) Using machine learning to detect cyberbullying. Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on.Vol. 2. IEEE.

[3] Chen, Ying, (2012) Detecting offensive language in social media to protect adolescent online safety. Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference onand 2012

[4] Fire M, Goldschmidt R, Elovici Y ,(2014) Online Social Networks: Threats and Solutions, Communications Surveys& Tutorials, IEEE.

[5] Munezero, Myriam, (2013) Antisocial behavior corpus for harmful language detection. Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on. IEEE.