Implementation of AI-Based Social Media for Vulgar Content Detector and Remover

Prof. Aarti Burghate[1], Poonam Bramhane[2], Pranjal Kuhikar[3], Pranjali Kanhekar[4], Ruchika Parate[5], Anjali Dongre[6]

Assistant Professor, Department of Information Technology[1]
Students, Department of Information Technology[2,3,4,5,6]
Nagpur Institute of Technology, Nagpur, Maharashtra, India

Abstract: *Social media is a web-based technology that makes it easier for a lot of people to communicate socially. Social media is used by billions of people worldwide to connect and share information. In this project, a social media network powered by AI will be developed for the purpose of identifying and eliminating offensive information. The platform uses machine learning and natural language processing to examine user-generated content and flag any posts or comments that employ offensive language or imagery. The technology is trained using both feedback from human moderators and a dataset of previously detected foul content. Additionally, the site has tools that users can use to report and remove inappropriate information. There are also systems in place to punish or ban individuals who frequently flout community rules. The platform also has a function for automatically detecting and removing vulgar content in real time utilizing language model and picture recognition technologies. By eliminating offensive information and encouraging constructive conversations, this project seeks to build a secure and welcoming online community for its members.*

Keywords: Vulgar Content, Instagram, Social Network

I. INTRODUCTION

The World Wide Web (WWW) and the Internet have made it simple for us to obtain information and knowledge. Social media has taken on a more significant part in people's lives as a result of advancements in internet technology. On social media, people can express their ideas and emotions. Internet-based social media is a form of communication. Users can converse and share information on social media networks. There are numerous social media platforms, including ones for instant messaging, video sharing, and photo sharing. browsing the many tweets, posts, images, and videos to find out what is happening in the lives of people they know. Social media has evolved into a platform for communication with friends and the exchange of viewpoints on global news and events. Social media is a fantastic method to pass the time similar to watching TV, especially with the expansion of picture and video media. Facebook, Twitter, Instagram, Tinder, YouTube, Snapchat, and a long list of other major social media platforms are just a few of the most used ones.

1.1 Instagram

What American computer Meta Platforms-owned photo and video sharing platform is Instagram? Users of the app can upload media that can be edited with filters, arranged by hashtags, and categorized by location. Public or pre-approved followers may share posts. Users can view trending material, like photos, follow other users to add their stuff to a personal feed, and browse other users' content by tag and location. Nowadays, social media is one of the most popular platforms due to the rising user base. Like other social media platforms, social media develops into a fantastic resource for information sharing.

1.2 Problem Statement

In social media, users upload their photos and videos. There is a section where other users can comment on those posts, including inappropriate, vulgar, bad, and negative comments. Users upload offensive photos that have an impact on people's behavior and imagination. On the internet, abusive language is defined by a wide range of terminology and norms, which can affect what is considered to be abusive language. The term "abuse" in the context of natural language

DOI: 10.48175/IJARSCT-8066

processing encompasses a variety of unfavorable phrases. "Any expression that is meant to disparage or offend a particular person or group," according to Mishra, is what the phrase refers to. In contrast to broad abuse, directed abuse concentrates on one particular person. Obscene picture identification is the process of locating explicit and pornographic content in images that have been previously extracted from a particular video file; it is the central component of a more comprehensive obscene- video filtering system. The skin colour and edge information of the image in question are tracked by existing obscene-image recognition techniques using information about image texture such RGB proportions, color-distribution histograms, and YIG. However, the obscenity level in low-quality UCC videos cannot be accurately determined using the currently available approaches. In order to establish whether an image passes the final obscenity test, this study provides an improved method that first uses Canny Edge to assess the fine grains of the image to determine whether it is of high or low quality. An arbitrarily chosen batch of photos was first subjected to the Canny Edge test to divide the batch into two groups based on the image- quality level in order to evaluate the effectiveness of this procedure.

II. LITERATURE SURVEY

AI-based social media strategies can be employed to prevent people from publishing hazardous content. Given the situation, this might lessen the need for online content moderation systems. The effective control of dangerous internet content is a difficult challenge for many reasons, although technology can help by limiting the amount of hazardous content posted for specific classes. AI is creating algorithms to identify harmful and offensive remarks, warn us about them, and flag them for deletion.

Farkhund Iqbal, Michael Motylinski, Kellyann Stamp, Mohammed Hussain, Andrew Marrington, Aine MacDermott "Using deep learning to find online bullies" Digital Investigation of the Year 2022, published by Forensic Science International

Although the idea of detecting illegal conduct online is not new, the ways in which it might happen are. In this evolving environment, technology and the proliferation of social media platforms and applications have a critical role to play. As a result, we see a growing issue with cyberbullying and "trolling"/toxicity on social media sites that share stories, postings, and memes. We discuss our research into using deep learning algorithms to identify "trolls" and harmful information disseminated on social media sites in this article. For better prediction abilities, we suggest a machine learning approach for the identification of harmful pictures based on integrated text content.

"Obscene Image Detection Algorithm Using High- and Low- Quality Images" appeared in the International Journal of Engineering and Industries in 2010 by Myoungbeom Chung and Daesik Jang.

Obscene picture identification is the technique of locating explicit and pornographic content in images that have been previously extracted from a particular video file; it is the central component of a more comprehensive obscene-video filtering system. The skin- color and edge information of the picture in question are tracked by existing obscene-image recognition techniques using information about image texture such RGB proportions, color- distribution histograms, and YIG. The approach suggested in this research is an improvement that uses Canny Edge to first analyse the fine grains of the picture to assess its quality before using to decide if it passes the final obscenity test.

Ashish Jhanwar1, Manoj K. Chinnakotla1, Jay Goyal1, Harish Yenala1, Deep learning for identifying offensive text content 2017's issue of International Journal of Data Science and Analytics

There are many online discussion forums available today that allow users to express, debate, and share their thoughts and ideas on a wide range of subjects. For instance, news websites, blogs, and social media sites like YouTube. Usually, comments are used to let people voice their opinions. It has frequently been noticed that user interactions in these forums may easily go off course and turn improper, including flinging insults or making harsh or inconsiderate remarks about specific people or groups/communities. Similar to this, certain virtual agents or bots have been observed to reply to users with offensive remarks. As a result, offensive messages and remarks are gradually losing their efficacy and becoming a threat online.

DOI: 10.48175/IJARSCT-8066

Hasanuzzaman, Mohammed, and Sidharth Mehra The International Journal of Data Science and Analytics will publish "Detection of Offensive Language in Social Media Posts" in 2020.
Social media abuse and inappropriate material posting have become major problems in recent years. Due to the massive popularity and use of social media platforms like Facebook and Twitter, this has led to several issues. The primary driving force behind this is the fact that our model will automate and speed up the identification of objectionable content that has been posted, making it easier to take the necessary measures to moderate these offending messages. For this research end behavior, we would be utilizing the openly accessible benchmark dataset OLID 2019 (Offensive Language Identification Dataset).

John Prakash and Neeraja M International Journal of Data Science and Research published "Detecting Malicious Posts in Social Networks Using Text Analysis" in 2015
The area of social media platforms like Facebook, Twitter, Flickr, and YouTube is here. More people utilize social networks than search engines when using the internet. To encourage direct connections with internet users, celebrities and businesses create social networking pages. Social media platforms mainly rely on people to share and obtain content. Information is quickly and efficiently shared throughout social networks. Social media networks do, however, at the same time become more open to various undesired and destructive hacker or spammer activities. It has been noted that Facebook pages have higher participation rates when it comes to the creation of harmful information.

III. PROPOSED WORK

In this web application, we will provide you with a social media platform that provides you with an online photo-sharing website or messaging platform, in this web application we will overcome with vulgar photos and videos. We'll also overcome negative comments.

3.1 Proposed Statement

A. Admin Module

Admin first registers and login himself on the website
On your website, users may add posts, leave comments, and engage in other activities by enabling user registration in social media. This post will demonstrate how to manage users and make it simple to facilitate user registration.

He has access to the registered users' comments and activity on our site. to quickly enable user registration and user management.
User feedback is data gathered directly from consumers or customers about how they felt while using a product, service, or website. Several approaches, such as surveys measuring customer satisfaction (CSAT) or net promoter score (NPS), are used to gather this data. UX designers, researchers, and marketers leverage user input and insight to enhance the user experience.

We will utilize the CNN algorithm to locate the issue, and with the aid of this algorithm, we will be able to erase all of the unnecessary and offensive remarks made on social media.
The foundation of several contemporary computer vision systems is comprised on convolutional neural networks, or CNNs. CNN is capable of completing the tasks of semantic segmentation, object identification, and image classification.
For deep learning algorithms, a CNN is a specific type of network architecture that is utilized for tasks like image recognition and pixel data processing. Although there are different kinds of neural networks in deep learning, CNNs are the preferred network design for identifying and recognizing objects.

3.2 Methodology

This web application was created using the Python Django Framework. We'll include a CNN algorithm with artificial intelligence.

The ability of artificial intelligence to close the gap between human and computer skills has been growing dramatically. Both professionals and amateurs focus on many facets of the field to achieve great results. The field of computer vision is one of several such disciplines. In recent times, CNNs have excelled in a number of NLP tasks. Our CNN simulation is motivated by (Kim, 2014). We provide a brand-new deep learning-based method for automatically recognizing.

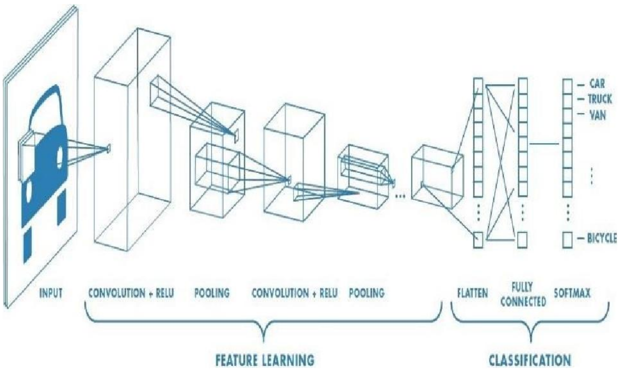

Figure 1: CNN Algorithm

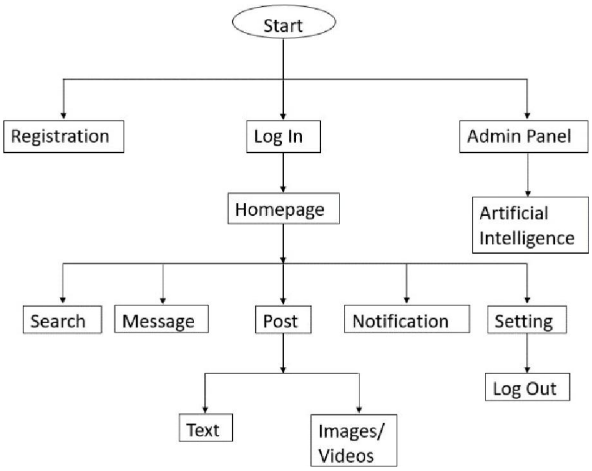

Figure 2: Flowchart

IV. RESULT ANALYSIS

Search system / Message / Post / Notification / Setting will be working properly.

- **Notifications:** Notifications means it gives some information about what's going on, on our website or some updates. When your Activity Status is on, people they follow and anyone you message will see when you were last active or currently active on your social media.
- **Search:** Hear you can find the people by their names and their IDs so you can make them your online friends. You can search on social media using keywords to find photos and videos, hashtags, accounts, and tags.
- **Message:** A message is a communication or statement conveyed from one person or group to another. If you call my house phone and I'm out running an errand, you'll be asked to "please leave a message after the beep." Generally transmitted verbally or in writing, a message can also be sent via a look or a gesture.
- **Post:** Picture description is a written caption that describes the essential information in an image. Picture descriptions can define photos and video anything containing visual information.
- **Algorithm will operate all the settings:** All nodes in a layer are completely linked to all nodes in the layer below them in a network that is fully connected. This generates a sophisticated model to investigate all CNN connections between nodes. The intricacy, however, comes at a great cost when training the network.

V. CONCLUSION

We are able to create a social media web application with all the features of Instagram. See how the public interacts with a social media app like Instagram in this article. In it, we offer a quick and simple step-by-step tutorial on how to create a social media app that may quickly attract plenty of active followers and successfully promote the reputation of your company online.

REFERENCES

[1]. Aghababaei, S., Makrehchi, M., 2017. Mining social media content for crime pre-diction. In: Proceedings - 2016 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE, pp. 526e531. https://doi.org/10.1109/WI.2016.0089 . WI 2016.

[2]. Al-Room, K., et al., 2021. ‘Drone forensics: a case study of digital forensic investigations conducted on common drone models’, international journal of digital crime and forensics. IGI Global 13 (1), 1e25. https://doi.org/10.4018/IJDCF.2021010101

[3]. Anand, M., Eswari, R., 2019. Classification of abusive comments in social media using deep learning. In: Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019. Institute of Electrical and Electronics Engineers Inc., pp. 974e977. https://doi.org/10.1109/ICCMC.2019.8819734

[4]. Arshad, H., Jantan, A., Omolara, E., 2019. Evidence collection and forensics on social networks: research challenges and directions. Digit. Invest. 28, 126e138. https://doi.org/10.1016/j.diin.2019.02.001 . Elsevier Ltd.

[5]. Bhattacharya, P., 2019. Social degeneration through social media: a study of the adverse impact of “memes”. In: ITT 2019 - Information Technology Trends: Emerging Technologies Blockchain and IoT. Institute of Electrical and Electronics Engineers Inc., pp. 44e46. https://doi.org/10.1109/ITT48889.2019.9075096 Boast, K., Harriss, L., 2016. Digital Forensics and Crime. POSTnote 520 March 2016.

DOI: 10.48175/IJARSCT-8066