IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |

Moderating Social Media Content: Exploring the Potential of Robotic Process Automation in Moderation of Social Media Content

Abhilash B N [1], Anusha M P[2], Bhavana K V [3], Harshitha M P[4], Prof. Naseer R[5], Prof. Ranjitha H S[6]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[1]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[2]

U.G. Student Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[3]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[4]

Assistant Professor,, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[5]

Assistant Professor,, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[6]

ABSTRACT: In this profound digital world, social media users have the liberty to express their views about any topic wherever they want. They might be objectionable. Explicit content such as profanity, spam, hate speech, nudity, offensive gestures, etc. Social media moderation plays a crucial role in monitoring user-generated content while protecting the online platform safe for users and brands. This process of moderation was earlier taken up manually by group of people or community and monitor thousands of posts that are being posted on the social media which is not humanly possible to go through each and every post so then AI came into picture, AI content moderation can help optimize the content moderation process by automatically analysing and classifying potentially harmful content, increasing the speed and effectiveness of the overall moderation procedure. Here we make use of Robotic Process Automation and automate these moderations which profoundly screens out the explicit content and also helps in reducing lots of time which involves both manual and AI algorithms to detect and report detrimental content.

KEYWORDS: Robotic Process Automation, Explicit content, social media moderation, detrimental content.

I. INTRODUCTION

Social media has had an extensive impact on the sector. It has linkedpeople globally, taking into consideration the unfolding of facts and the formation of relationships. Social media is crucialbecause it has the capacity to distribute information to the completeglobal. It creates connections, stocksstatistics, and facilitateshumans to locate what they'reseeking out. Social media alsopermitscorporations to understand their audiencehigher and supplypreciousproducts or content material . However, social media also can be addictive and featurebadconsequences on mentalfitness. Universal social media has itsadvantagesand drawbacks, and its effect on society is complex and

e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |

multifaceted. Social media could have a damagingeffect on mentalfitnessbecause of the poorcontentit is able topromote. Heavy social media use has been related to an elevatedrisk of melancholy, tension, loneliness, self-damage, or even suicidal mind. Social media also cancauseemotions of inadequacy about one's life or look. Social media use has been tied to reduced, disrupted, and not on time sleep, that isassociated withmelancholy and reminiscence loss. Social media structures are designed to be addictive and are related totension, despair, and evenaddiction. If social media use becomesa substitute for offline social interplay, it could adversely have an effect onmentalhealth.

So, the principle root cause of social media being this pessimistic for some set of users is that the sorts of posts being publishedbyusers which succeedat some stage in the net and it ishardto manipulate which set of customers are eating this statistics and being affected these posts or factspostedat the social media platformsmay additionallyownthingsrelated to violence, nudity, fauxinformationand so on. This trouble is being addressed in manyfeasibleways. I find it irresistibleto begin with guideassessment of posts once they are posted on social media platforms, where this solution is very unrealistic and of course not a feasible one as manual evaluation would be very time consuming as well as non-productive and not accurate. Then came the next solution where they are using AI algorithms to moderate content .these are also to some extent feasible as they classify posts based on the previous training.but this can't completely serve purpose as this classification is based on pretrained data and this process includes manually giving posts as input to the algorithm which inturn consumes a lot of time.

II. RELATED WORK

Social Media, Content Moderation, and Technology [1]. Authors: Yi Liu, T. Pinar Yildirim, Z. John Zhan . In this paper, we develop a theoretical model to address all these questions. In our model, a plat-form allows users to post and share content with others, and earns revenue either through advertising or subscription fees. Users enjoy the ability to express their opinions, and read others' content, from which they may or may not obtain positive utility depending on their own preferences and also on how extreme the content on the platform is. A platform moderates online content to maximize its revenue.

Detection and moderation of detrimental content on social media platforms [2]. Authors: Vaishali U. Gongane & Mousami V. Munot. The literature review shows that there is a massive amount of research explored in the detection methods of various forms of detrimental content. From a theoretical point of view, reported articles have focused more on the various aspects involved in terms of manual method of moderation and challenges that the AI based methods should address. There are less research articles that focus on fully automated moderation techniques of detrimental content on social media platforms.

Social media content moderation: Six opportunities for feminist intervention [3]. Authors: Gerrard, Y. These highlight moments when social media content is removed or hidden from view, but actions like removal are the ends points of human-algorithmic systems. This short essay explains why we get to those end points, outlining six stages of the content moderation process that are most fallible to human intervention (and therefore bias, subjectivity, intolerance). Somewhat optimistically, it also explains how these ideology-laden spaces might also be opportunities for feminist intervention. In short, they are the spaces we need to target if we want to enact change in the system.

Platform's content moderation strategy under imperfect technology [4]. In reality, an accurate technology for content moderation is still many years into the future (Singh, 2020). As Mark Zuckerberg commented, "over a five-to-ten-year period we will have AI tools that can get into some of the linguistic nuances of different types of content to be more accurate, to be flagging things to our systems, but today we're just not there on that... There's a higher error rate than I'm happy with" (Gershgorn, 2020). In a well-publicized example, in the days leading up to 4th of July, 2018, Facebook's algorithm for "hate speech detection" flagged down and removed a post of the Declaration of Independence because of paragraphs 27-31, which include the phrase "merciless Indian savages" (Sandler, 2018). The existence of imperfect technology raises a number of questions about the practice and management of content moderation. First, if technology has a "higher error rate" than a platform is "happy with," how should a platform employ the technology given the choice of its revenue model? In this regard, a related question is whether a platform has the incentive to

e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |

embrace an inaccurate technology to do content moderation. Second, how can a platform best manage its content moderation to achieve its profit objectives. Given that a platform's primary objective is to maximize its profit, could content moderation with imperfect technology lead to a higher extremeness index for the platform. Finally, if today's technology is "just not there," and there is "a higher error rate," what kind of a platform has the most incentives to improve it or not improve it.In this section, we address all those questions by extending our model to incorporate imperfect technology for content moderation.

III. METHODOLOGY

We developed an RPA (Robotic Process Automation) robot that can be deployed on an Ui path component called Orchestrator. The Orchestrator is a web-based application. It provides options to deploy, monitor, schedule, and control software bots and processes. It is a centralized platform used to control/manage all software bots. Our robot will be triggered at the time of uploading posts in social media. When a post gets uploaded in social media, it will be stored in a folder, the robot will be waiting in the folder for a new post, then it copies the content from that folder to a Local folder. We are copying the post to a local folder and our bot runs in the local folder, because if the bot runs on the social media folder it will affect the efficiency and speed of the website. The bot scans individually through the posts and text to check for any possible explicit content.

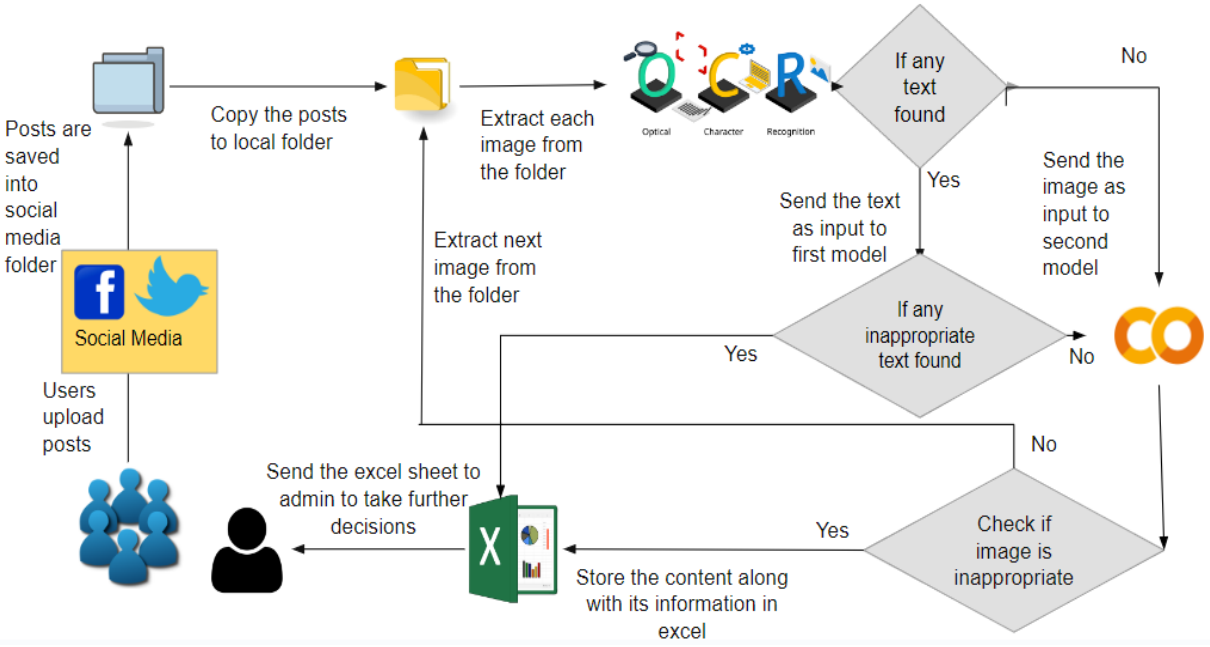


Fig 3.1: Workflow

Administrator does not want the bot to be working directly on the social media website folder. Once the bot detects that new files have been uploaded, it reads each file and sends those images to the algorithm. We use OCR to extract text from the images.image contains any text, then it is sent to the first algorithm to check for the inappropriateness, if there are no texts found in the image then sent to the second algorithm. Both the algorithms will be running on the google colab, the first algorithm uses Natural language processing algorithm which uses NLTK package to detect explicit content in text. and the second algorithm uses Google Cloud Vision API. These algorithms check for the appropriateness of the images uploaded. If the images are found to be inappropriate, then it is appended to an Excel log and this log is passed to the administrator

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

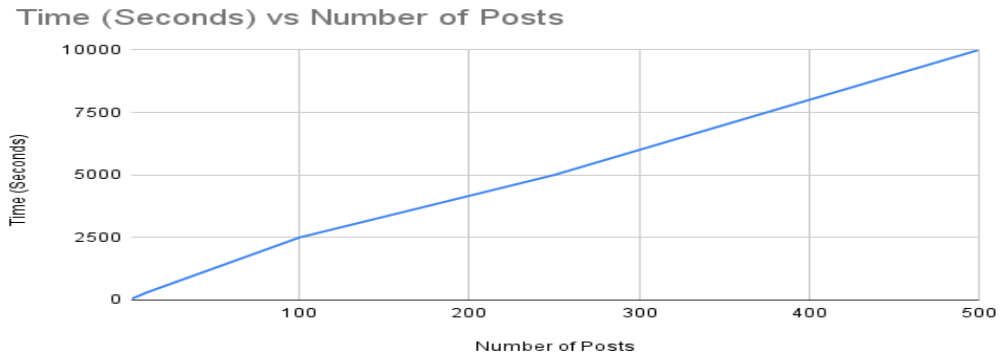
e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |

IV.EXPERIMENTAL RESULTS

As per the experimentation carried out on the project as mentioned in the methodology we have measured the time taken by the algorithm to classify the posts ,and this time is varying based on the number of posts it is trying to classify. We have represented the results of analysis made about execution time as below.



.4.1: Analysis of execution time

Below is the equation we have used to find the accuracy of the posts being classified.

$$\text{Accuracy} = \frac{\text{Correctly Classified posts}}{\text{Total number of posts}}$$

The graph below shows the accuracy of classification

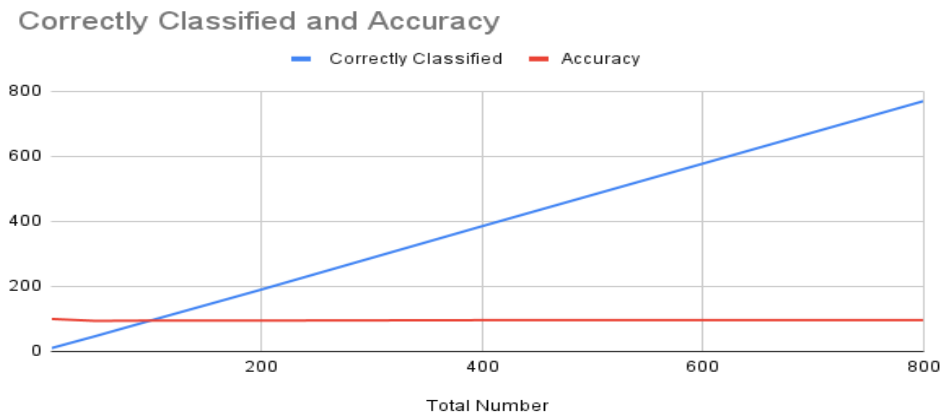


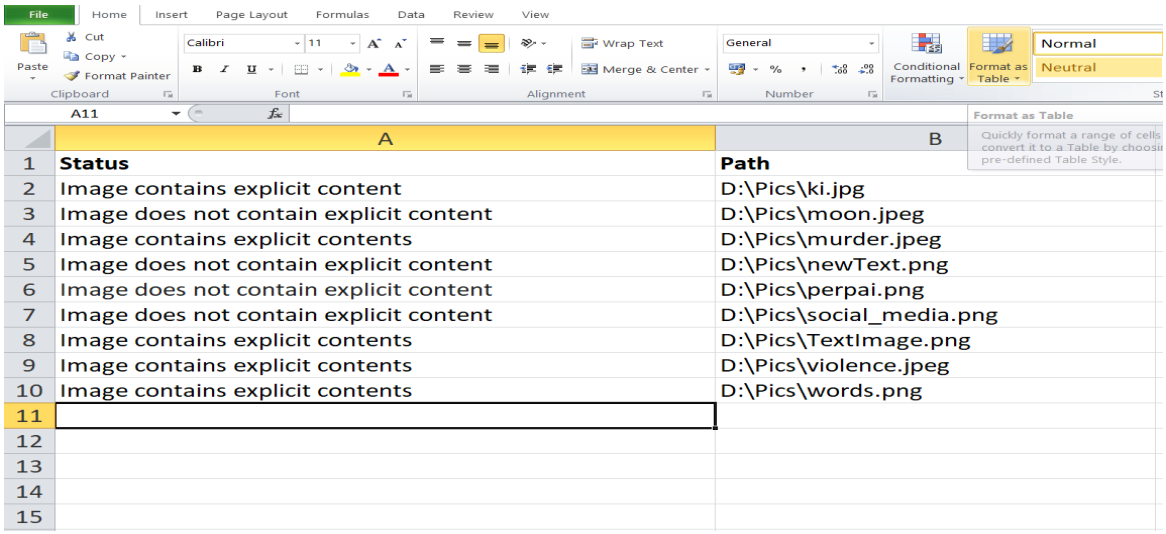
Fig 4.2: Calculating Accuracy

After the implementation as per the methodology that we designed, the following is the snapshot of the output that we are going to send to the admin of social media in order to take action over the explicit content being posted.

e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |



	A	B
1	**Status**	**Path**
2	Image contains explicit content	D:\Pics\ki.jpg
3	Image does not contain explicit content	D:\Pics\moon.jpeg
4	Image contains explicit contents	D:\Pics\murder.jpeg
5	Image does not contain explicit content	D:\Pics\newText.png
6	Image does not contain explicit content	D:\Pics\perpai.png
7	Image does not contain explicit content	D:\Pics\social_media.png
8	Image contains explicit contents	D:\Pics\TextImage.png
9	Image contains explicit contents	D:\Pics\violence.jpeg
10	Image contains explicit contents	D:\Pics\words.png
11		
12		
13		
14		
15		

Fig 4.3: Excel sheet showing the results of detection of explicit content

V. CONCLUSION

Content moderation is a very important process to be done on the social media platforms as the explicit content being posted on the platform is increasing down the lane. This will need ample research and good understanding of social media content posted by users. while designing an automated content moderation system as much as possible.

The detrimental content posted on social media has already caused damage to society. Present systems focus on moderating it manually and removing it after the damage has already been done. Our system automates this process and ensures the moderation before any harm is done. Here as per our methodology using robotic process automation, we can automate the detection of explicit content from posts which are being posted by users. After detection we further report the list of posts containing explicit content to the admin with an excel sheet.as we use automation and Ai hand in hand this could be a better moderation technique. But to the best interest of mankind and humanity, researchers need to think beyond moderating the content and going step further to prevent it wherein there is some flagging assigned to a user and after a threshold is decided on a number of inappropriate posts. So, the researchers need to think beyond the obvious of only moderating or only restricting their research to moderation of content, but prevention of such cases will actually serve as a boon to social media.

REFERENCES

[1]. Social media content moderation: six opportunities for feminist intervention. https://eprints.whiterose.ac.uk/162434/3/feminist-media-studies-C%2BC-GERRARD-%28draft2%29.pdf

[2]. Detection and moderation of detrimental content on social media platforms https://link.springer.com/article/10.1007/s13278-022-00951-3#author-information

[3]. Social media content moderation: Six opportunities for feminist intervention. Authors: Gerrard, Y. https://eprints.whiterose.ac.uk/162434/3/feminist-media-studies-C%2BC-GERRARD-%28draft2%29.pdf

[4]. Platform's content moderation strategy under imperfect technology

[5]. Robotic process automation https://en.wikipedia.org/wiki/Robotic_process_automation

[6]. What is robotic process automation? https://www.techtarget.com/searchcio/definition/RPA

e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204335 |

[7]. Robotic Process Automation (RPA), Automation software to end repetitive tasks https://www.uipath.com/rpa/robotic-process-automation

[8]. Automate more with the UiPath Platform https://www.uipath.com/

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details